面向隐马尔可夫特征的数据质量控制模型

周金陵 刁兴春 周 星 曹建军

(解放军理工大学 指挥信息系统学院，江苏 南京 210007)

摘 要：为了在进行数据质量控制时，考虑数据的具体特征以提高数据质量，分析了当前基于贝叶斯网络进行数据质量控制的模型、改进模型及其效果。针对字段之间存在“隐马尔可夫”关系的数据，提出了一种面向隐马尔可夫特征的数据质量控制模型，利用贝叶斯网络结构算法确定字段之间是否存在的“空间”上的关系，并且利用隐马尔可夫模型的 Baum-Welch 算法学习字段之间的生成概率矩阵（反映字段之间的依赖关系）和记录之间转移概率矩阵（反映记录之间的时间依赖关系），作为推理的依据。仿真实验表明，将“空间”和“时间”上的依赖关系有机结合起来预测或校验数据，对于存在“隐马尔可夫”关系的数据质量控制的提升效果显著，验证了模型的有效性。

关键词：数据质量；隐马尔可夫模型；贝叶斯置信网；时间依赖

A Hidden Markov Feature-Oriented Data Quality Control Model

Zhou Jin-ling, Diao Xing-chun, Zhou Xing, Cao Jian-jun

(College of Command Information Systems, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: Data quality control based on data mining technology is currently a hot topic in the research field. The paper analyzes data control models based on Bayesian Belief Network, improved models and their effects. The paper suggests that data features should be considered to build the effective model, a data quality control model is proposed to deal with the data including Hidden Markov feature in the paper. The model firstly builds the Bayesian network to determine the relationship between fields, then learns the generated probability matrix between fields (reflects the spatial dependency) and the transition probability matrix among the records(reflects the time dependency) as the foundation of reasoning. The model borrows the characteristic of Hidden Markov model, integrating the time and spatial dependency, improved data quality with Hidden Markov feature, and without loss of generality in the common data quality control.

Keywords: Data Quality, Hidden Markov Model, Bayesian Belief Network, Time Dependent

1 引言

随着大数据时代的来临，数据挖掘技术已经被广泛应用到各个领域进行数据分析以辅助决策。大多数的数据分析认为数据是“准确”、“干净”的[1]。而实际中，数据大多会存在不同程度的质量问题（缺失、重复记录、逻辑错误甚至伪造数据），严重的数据质量问题会给推理带来大的偏差，影响决策，因此，数据质量控制逐步成为数据科学领域研究的重要课题。

近年来，利用数据之间的统计依赖关系进行数据质量控制、提高数据的准确性成为研究热点，贝叶斯置信网理论是其中的典型应用之一。文献[2]通过表单的历史录入数据训练出反映字段之间的贝叶斯置信网，利用该贝叶斯置信网对随后的数据录入进行推理性的预测和数据校验，以达到提高数据质量的目的。但是，该文献[2]仅仅通过字段之间“横向”的依赖关系对数据进行校验，并未考虑了单个字段“纵向”取值之间的依赖关系。文献[3]在文献[2]的基础上，增加了这样“纵向”的依赖关系的校验——利用一个简单的不超过三元的马尔可夫链，将记录之间的字段值的关系联系起来，进一步提升了表单的数据质量。除了上述比较典型的阐述模型

基金项目：国家自然科学基金资助项目（61371196）；中国博士后科学基金特别资助项目（201003797）；解放军理工大学预研基金项目（20110604）

作者简介：周金陵，男，1987 年生，博士研究生，主要研究方向为数据工程；刁兴春，男，研究员，博士生导师；周星，男，1988 年生，博士研究生；曹建军，男，博士。

联系方式：13770542301，E-mail：zjl.wb@163.com。

的文献之外，还有一些利用贝叶斯网络[4]和关系马尔可夫模型[5, 6]处理缺失数据、提高数据质量的文献：比如，文献[4]利用简单的贝叶斯方法估计“枚举型数据”的缺失值后，分别根据最大化概率和后验分布，提供了两种可选且有效的填充缺失值的方法，文献[5]和[6]则将充分考虑属性间的关联性，将动态属性选择方法与关系马尔可夫模型结合，最大限度地利用完整数据的信息，提高了利用关系马尔可夫模型对枚举型数据的估计能力。这些研究都是以贝叶斯理论和马尔可夫理论为基础提高数据质量的实例，丰富了这类理论在数据质量控制中的内容。

通过分析相关研究可以发现现有工作存在的一些不足：第一，尽管利用贝叶斯置信网描述表单形式的数据字段之间的关系具有普适性，但是如果字段之间的统计依赖程度不深（即不能形成较完善的贝叶斯置信网，存在大量孤立字段），这种建立具有普适性的贝叶斯置信网方法在数据质量控制的效果上会大打折扣；第二，文献[3]加入的马尔可夫链有效地提高了相应字段的数据质量，但是这种对两种方法进行机械的加权的方式割裂了字段之间可能存在的内在联系，而且对于不具马尔可夫性的字段的质量控制效果有待验证。

现实中的数据不可能脱离特定的领域而独立存在，其中有一类数据含明显的时间依赖关系，两个字段之间则可能构成隐马尔可夫关系。比如病人病例中的各项体征数据、气象数据的风速、气温等，都具备上述特征。事实上，文献[3]所进行的研究和实验就是建立在此类数据之上引入马尔可夫链的。

基于现有研究的局限，考虑特定领域数据的特点，本文针对具有隐马尔可夫特征的关系型表单数据，提出了一种数据质量控制模型。该模型面向具备隐马尔可夫特征的数据建立，对于具备该特征的数据具有良好的质量控制效果，而且对于普通数据的质量控制效果也不失一般性。

2 相关工作

贝叶斯置信网，属于概率图模型的理论范畴[7]。一个贝叶斯置信网$\boldsymbol{B}$由网络结构$\boldsymbol{S}$和条件概率集合$\boldsymbol{P}$构成，即$\boldsymbol{B}=(\boldsymbol{S}, \boldsymbol{P})$。其中，$\boldsymbol{S}$是有向无环图：图中的节点对应随机变量，节点之间的有向边代表相应的随机变量存在概率依赖，如图1所示。图中代表随机变量的节点X_i，可以是任何领域内实体的抽象，如身高、体重、气温、湿度等。每一个随机变量（节点）拥有一个条件概率表（CPT，Conditional Probability Table），CPT反映了当前节点与其父节点的条件概率关系，根节点的CPT描述了随机变量的先验概率。

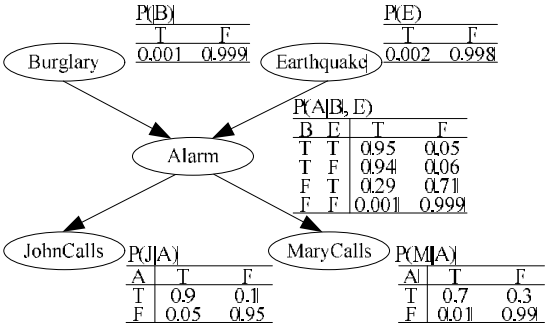

图1 贝叶斯网络示意图

对于每一个节点X_i和它的父节点集$\boldsymbol{\pi}_i$，满足如下公式

$$P(X_1, X_2, \cdots X_n) = \prod_{i=1}^{n} P(X_i \mid \pi_i) \qquad (1)$$

该公式为多变量全概率公式，蕴含了条件独立性假设——可以使用“D-分隔”(D-Separation)[8, 9]进行解读。利用该公式可以进行丰富的条件概率推理。

基于贝叶斯置信网进行推理的一般性步骤为：

(1) 贝叶斯置信网结构学习：即找到和样本数据集匹配最佳的贝叶斯置信网结构图$\boldsymbol{S}$，目前相关学习算法主要分为两类[10]：基于搜索和评分的方法和基于独立性测试的方法；

(2) 利用最大似然估计算法、贝叶斯估计算法等学习贝叶斯置信网参数，即每个节点的CPT构成的集合$\boldsymbol{P}$；

(3) 基于上述全概率公式，利用构建好的$\boldsymbol{B}=(\boldsymbol{S}, \boldsymbol{P})$进行推理（预测、校验、填充等）。

利用贝叶斯置信网对关系型表单进行数据质量控制，首先将关系型数据表的元素抽象为贝叶斯置信网，如图2所示：每一个字段对应贝叶斯置信网中的一个变量节点，该字段下的值$x_{ij(j=1,2,\cdots)}$是字段变量X_i的取值，每一条记录是一组数据样本。

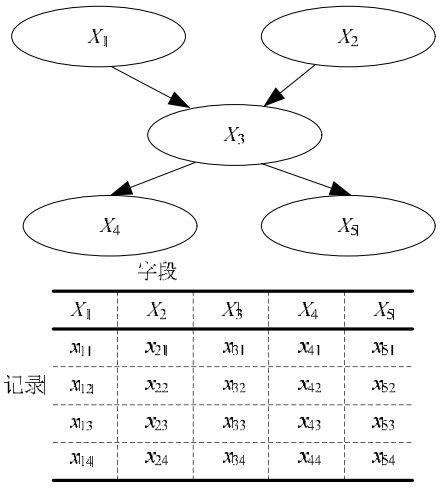

图2 基于表单数据建立BN模型

图3给出了基于贝叶斯置信网的数据质量控制框架的一般思路，也就是文献[2]中作者给出的控制框架的基本结构。该框架给出了数据质量控制的一般步骤：

(1) 将数据样本及字段名进行预处理，利用相关的算法构建数据的贝叶斯置信网结构图 $\boldsymbol{S}$；

(2) 基于上述训练数据学习贝叶斯置信网的参数，即CPT构成的集合$\boldsymbol{P}$；

(3) 基于公式(1)及其变形进行相关推理，预测即将录入的数据或校验已录入的数据，从而达到保证数据质量的目的。

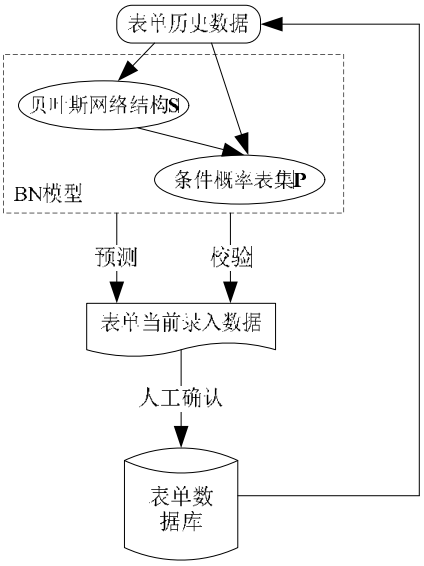

图3 基于BN的数据质量控制模型

文献[2]的作者在他的另一篇文献中[11]，设计了一组自适应的智能人机接口用以提高数据输入的准确性和效率。比如，根据历史数据，可能性很高的值被设为默认值；根据被选的概率，对答案进行动态的重排和高亮显示；对判断疑似错误的录入向用户发出警告信息等等。实验证明，这些带输入警告和提示的控件对数据录入准确性的提高大有裨益。通过对智能提示单选控件的使用实验发现，错选率下降了54-78%，而且在输入时间消耗和准确性上，具有很好的“性价比”。

上述框架简单而实用，经常会被借鉴到基于贝叶斯置信网的数据质量控制中。贝叶斯置信网的推理主要是基于字段变量之间依赖关系的“横向”推理，即利用同一条纪录的不同字段之间的关联进行推理，一般情况下能够很好的进行数据的预测和校验，满足数据质量控制的要求。但是当字段之间的关联性比较弱的情况下，这样的“横向”推理的效果就显得一般。为此，相关的研究进一步考虑了同一字段，记录之间的地“纵向”依赖关系，改进了上述模型[3]，如图4所示。

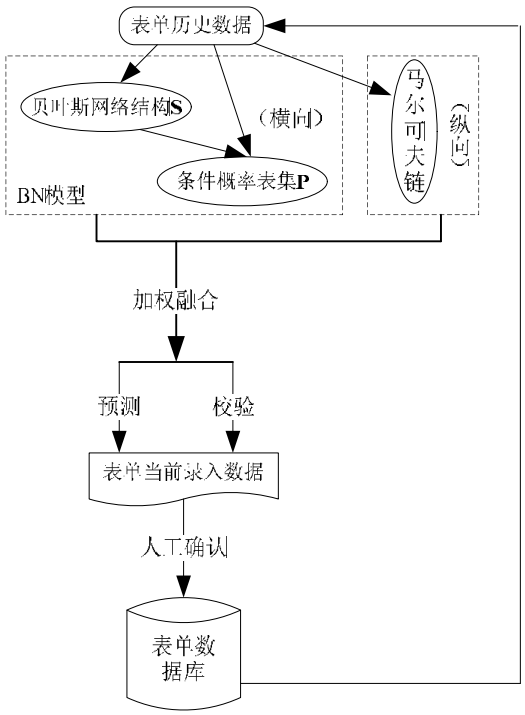

图4 基于BN和Markov Chain的数据质量控制模型

对于图2表单中的某个字段X_i的值x_{ij}，利用图4中的框架根据该字段的历史值预测其出现的“纵向”概率$p_M(x_{ij})$为：

$$p_M(x_{ij}) = \alpha_3 \times p_m(x_{ij} \mid x_{i\,j-1}, x_{i\,j-2}) + \alpha_2 \times p_m(x_{ij} \mid x_{i\,j-1}) + \alpha_1 \times p_m(x_{ij}) \quad (2)$$

公式(2)中的3项分别为二阶马尔可夫概率、一阶马尔可夫概率和先验概率，其中的加权值用来依据重要性的程度给各项分配权重，同时进行归一化处理，即

$$\alpha_3 + \alpha_2 + \alpha_1 = 1 \quad (3)$$

因此，x_{ij}的综合预测概率为

$$p(x_{ij}) = \beta \times p_M(x_{ij}) + \gamma \times p_B(x_{ij} \mid x_{kj}, x_{lj}, \cdots) \quad (4)$$

$$\beta + \gamma = 1 \quad (5)$$

公式(4)中的 $p_B(x_{ij} \mid x_{kj}, x_{lj}, \cdots)$ 为上述一般框架中，利用贝叶斯置信网预测的概率，其中为 $x_{kj}, x_{lj}, \cdots$ 为 x_{ij} 的父节点，该概率是x_{ij}的CPT参数。

当推理的当前字段存在马尔可夫性时，图4中改进的模型一般能够增加推理的准确性（文献[3]的实验证明准确率可以提高约7%左右）。

3 面向隐马尔可夫特征的数据质量控制模型

文献[3]中针对数据质量控制的研究已经考虑到数据的特点，比如数据可能包含时间依赖关系，但是这种对两种方法进行机械的加权方式割裂了字段之间可能存在的内在联系，实际字段值的随机情况也往往没有文献[3]中描述的那么直观明了，而是更加复杂。

3.1 数据质量控制模型

通常数据中所能观察到的字段信息并不是与Markov链的可能状态一一对应的，而是通过一组概率分布与Markov链的各个状态（另一个字段的状态值）相联系。这样构成一个双重的随机过程，其中一个用Markov随机过程描述字段值之间的转移概率，另一个随机过程用来表示Markov随机过程描述的字段与另一个字段可能出现的观测值之间的统计概率对应关系，这种模型被称为“隐马尔可夫模型”（HMM，Hidden Markov Model）[12, 13]。比如在医学数据中，患者的体重观测值不仅和BMI（身体质量指数）有很强的关联，而且能够反映病患的BMI的变化；气象数据中，风速数据的观测值不仅能够反映气温的可能值，并能够反映气温数据的变化规律。这种类型的数据被称为具备“隐马尔可夫特征”。为了更好地控制具备这类特征的数据质量，需要在构建控制模型时考虑这种特征。

这里利用 HMM 对数据字段之间的概率关系进行描述。假设在上图 2 中，希望通过某病患健康报告 X_3 的值（比如体重值）推理出 X_2 的信息（比如 BMI 值），已知 X_2 包含 n 个状态值，并且根据先验知识可以得到这 n 个状态值的先验概率向量$\boldsymbol{\pi}$和状态之间的转移概率矩阵 **TRANS**。$\boldsymbol{\pi}$和 **TRANS** 可以通过历史数据得出。X_3 作为 X_2 的一个观测值，一共有 m 个值状态。这 m 个值与 X_2 的 n 个状态关系为一个观测矩阵（生成概率矩阵）：

$$\mathbf{EMIS} = (q_{lk}) = \begin{bmatrix} q_{11}q_{12}\cdots q_{1n} \\ q_{21}q_{22}\cdots q_{2n} \\ \cdots\cdots \\ q_{n1}q_{n2}\cdots q_{nn} \end{bmatrix}$$

q_{lk} 表示 X_2 为 l 值时，观测到 X_3 为 k 值的概率。

以上便建立了一个隐马尔可夫模型。当 X_3 观测到输入序列为 $X_{31}, X_{32}, \ldots$时，通过以上所建立的模型便可概率地计算出 X_2 可能的字段序列。

简而言之，隐马尔可夫模型就是基于马尔可夫假设和独立输出假设，在已知观测字段的连续记录值 $x_{j1}, x_{j2}, x_{j3}, \cdots$ 的情况下，求得令条件概率 $P(x_{i1}, x_{i2}, x_{i3}, \cdots \mid x_{j1}, x_{j2}, x_{j3}, \cdots)$ 达到最大值的那个字段连续记录值，即

$$x_{i1}, x_{i2}, x_{i3}, \cdots = \underset{all\ x_{i1}, x_{i2}, x_{i3}, \cdots}{\mathrm{ArgMax}} P(x_{i1}, x_{i2}, x_{i3}, \cdots \mid x_{j1}, x_{j2}, x_{j3}, \cdots) \quad (6)$$

通过推导（步骤略），可以转化为计算某个特定的状态序列（在这里是连续的记录字段值）$x_{i1}, x_{i2}, x_{i3}, \cdots$ 产生出输出符号 $x_{j1}, x_{j2}, x_{j3}, \cdots$ 的概率，即

$$P(x_{i1},x_{i2},x_{i3},\cdots x_{j1},x_{j2},x_{j3},\cdots)$$
$$=\prod_t P(x_{it}\mid x_{i(t-1)})\cdot P(x_{jt}\mid x_{it}) \tag{7}$$

这里的马尔可夫假设和独立输出假设是指

$$P(x_{j1},x_{j2},x_{j3},\cdots\mid x_{i1},x_{i2},x_{i3},\cdots)$$
$$=\prod_t P(x_{jt}\mid x_{it}) \tag{8}$$

$$P(x_{i1},x_{i2},x_{i3},\cdots)=\prod_t P(x_{it}\mid x_{i(t-1)}) \tag{9}$$

如图5是上述关系数据字段的HMM组成图。

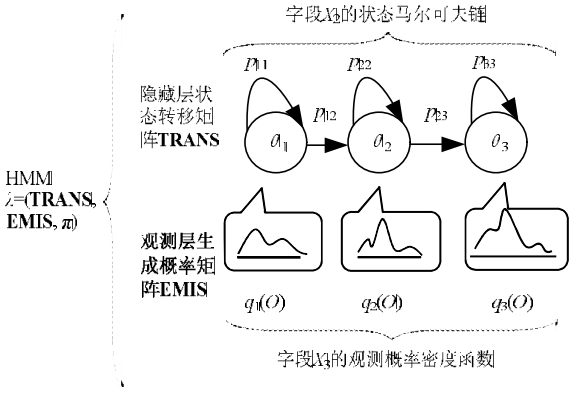

图5 关系数据字段的HMM组成图

基于贝叶斯置信网和隐马尔可夫理论的数据质量控制模型如图 6 所示，根据该框架建立模型的步骤为：

(1) 将数据样本及字段名进行预处理，利用相关的算法构建数据的贝叶斯置信网结构图 ***S***；

(2) 找出贝叶斯置信网结构图中具有因果关系的字段“变量对”，据HMM理论，一般将因变量设为“隐藏变量”，根据训练数据计算其转移概率矩阵**TRANS**，果变量设为“观测变量”，计算其生成概率矩阵**EMIS**；

(3) 联合**TRANS**和**EMIS**进行推理，预测即将录入的数据和校验已录入的数据，从而达到保证数据质量的目的。

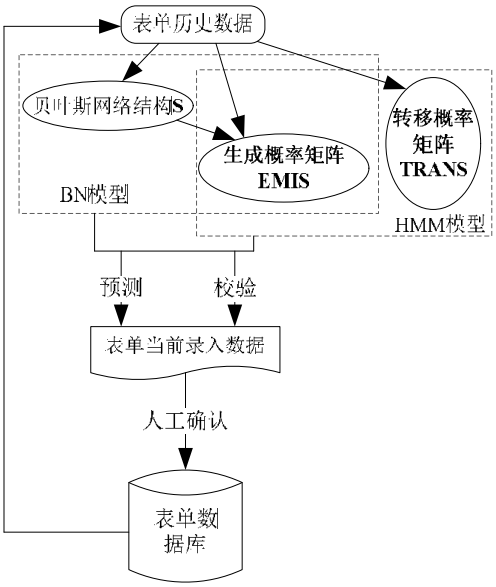

图 6 基于 BN 和 HMM 的数据质量控制模型

3.2 模型中的关键理论问题

为了具体说明问题，就图 2 的表单应用说明模型中的关键理论。

图 3 基本的控制框架中，按照步骤需要先行建立网络结构 ***S***，***S*** 反映了由历史数据训练出的字段变量之间最可能的因果关系，比如 X_3 有两个因变量 X_1 和 X_2，确定了这样的因果关系后，在接下来的参数学习中从历史数据中学得的参数为联合条件概率，即

$$P(X_3\mid X_1,X_2)=\frac{P(X_3,X_1,X_2)}{P(X_1,X_2)} \tag{10}$$

对于 X_3 而言，其 CPT 参数是一个 $(|X_1|\times|X_2|)\times|X_3|$ 维的矩阵（其中 $|X_i|$ 为变量 X_i 取值的个数）。假设 X_2 值未知(缺失、尚未填写)，只能依据 X_1 值来推理 X_3 的值，则需要对 X_2 的全部情况求和，即：

$$P(X_3\mid X_1)=\sum_{X_2}P(X_3\mid X_1,X_2) \tag{11}$$

本文所提的模型一方面简化了贝叶斯置信网中的参数学习的步骤，使用 HMM 中的 **EMIS** 矩阵(反映两个字段之间的生成概率关系)代替了 CPT 参数；另一方面使用 HMM 中的 **TRANS** 矩阵代替马尔可夫转移概率。贝叶斯置信网中变量的因果关系可能

是多元的，即 X_3 有两个因变量 X_1 和 X_2（而不是一个），所以一般在条件充足的情况下，使用联合条件依赖全面描述这 3 个变量的依赖关系，即上述式(10)。而 HMM 是两两变量之间生成关系，即 $P(X_3|\ X_1)$和 $P(X_3|\ X_2)$，这与 $P(X_3|\ X_1, X_2)$没有必然的联系。因此，在 X_1 和 X_2 都已知的条件下，为了更加准确的反映推理结果，同时简化推理的复杂性，使用下式来近似代替 $P(X_3|\ X_1, X_2)$：

$$\begin{aligned} &P(X_3 \mid X_1, X_2) \\ &\approx \mu P(X_3 \mid X_1) + \nu P(X_2 \mid X_1) \end{aligned} \tag{12}$$

$$\mu + \nu = 1 \tag{13}$$

式(12)中的两项都是直接来自**EMIS**矩阵，当条件不充分的情况下（比如仅仅已知X_1的值），推理直接使用其中一项即可。

4 仿真实验

4.1 数据集的选择和实验环境

为了对含 HMM 特征的数据和一般数据的质量控制的效果进行比较分析，利用 Matlab 的贝叶斯网络工具箱[14]的 sample_bnet()函数和隐马尔可夫模型工具箱[15]的 hmmgenerate()函数生成了两组同样数量训练数据，并组合到一个数据集中。为了具有可比性，同样利用文献[2]使用的贝叶斯网络结构学习的工具 Banjo[16]学习字段之间的关系（图 7）。从图 7 中可以看出，Banjo 能够将来自两个数据集的字段有效的分开（左右两部分），并且比较准确构建了对各个数据集字段内部的因果关系。

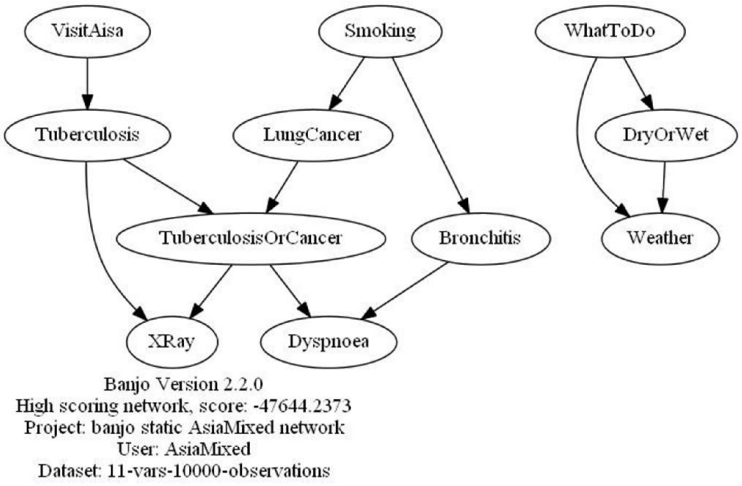

图 7 利用 Banjo 学习出的数据贝叶斯网络结构

4.2 实验结果

基于上述数据集和学习出的贝叶斯网络结构，将不同数目的数据集按照 8:2 的比例分为训练数据和测试数据，使用本文提出的模型进一步学习模型的参数，即描述两两变量关系的生成矩阵 **EMIS** 和记录之间转移概率矩阵 **TRANS**。利用学习好的参数在测试数据集上验证推理的准确率。

为了比较含 HMM 特征的数据和一般数据的质量控制的效果，分别测试了 Smoking 字段值和 WhatToDo 字段的预测效果，并且同现有的模型的平均准确率进行了比较。

从性能对比中（图 8）可以看出：

(1) 当数据量达到千条以上时，数据推理的准确性收敛到相对稳定的水平；当数据量比较小时，推理的效果不能保证。
(2) 对于一般数据推理的准确程度与基于贝叶斯置信网的模型的能力相当，约 65%左右。
(3) 对于含有时间依赖关系、具有隐马尔可夫特征的数据值推理的准确程度可以达到 80%以上，优于单纯的对贝叶斯理论和马尔可夫链的加权方法（70%左右）。

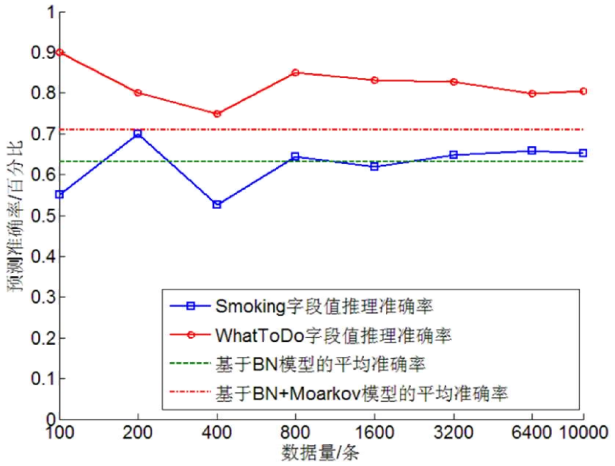

图 8 不同模型的数据质量控制性能对比

因为本文所阐述的用于保证、提高数据质量的方法——预测、自动填充、校正等，本质上都是借助数据挖掘中的推理机制，所以本文（包括仿真实验）并没有区分具体的数据质量控制的手段。

5 总结

使用数据挖掘技术检测、填充和校正数据，能够有效地提高数据质量。本文研究了一般的基于贝叶斯置信网的数据质量控制模型和相关的改进模型，提出了结合数据特征进行数据质量控制的方法，并将该方法抽象成类似的模型。结合前人的相关工作，研究得到如下结论：

(1) 与单纯的加权方法相比，利用 HMM 的生成矩阵和转移矩阵，将“空间”上字段之间的“横向”依赖和“时间”上记录之间的“纵向”依赖有机结合起来，推理更具合理性；

(2) 现有模型下的数据质量进一步提升，针对特定特点的数据建立的模型同样可以不失通用性；

(3) 在一定程度上简化了模型参数学习的复杂度，有利于提高数据质量控制的效率，日后需要对模型的“性价比”进行量化的验证。

参考文献：

[1] Dasu T., Johnson T. Exploratory data mining and data cleaning[M]. John Wiley & Sons, 2003.

[2] Kuang Chen, Harr Chen, Neil Conway, et al. Usher: Improving data quality with dynamic forms[J]. IEEE Transactions on Knowledge and Data Engineering. 2011, 23(8): 1138-1153.

[3] Yuan Ling, Yuan An, Mengwen Liu, et al. An Error Detecting and Tagging Framework for Reducing Data Entry Errors in Electronic Medical Records(EMR) System[C]. International Conference on Bioinformatics and Biomedicine, 2013: 249-254.

[4] Li Xiao-Bai. A Bayesian Approach for Estimating and Replacing Missing Categorical Data[J]. ACM Journal of Data and Information Quality. 2009, 1(1): 1-11.

[5] Cao Jianjun, Diao Xingchun, Xu Yongping, et al. An Approach Using Relational Markov Model for Estimating and Replacing Missing Categorical Data[C]. The 18th International Conference on Information Quality(ICIQ 2013), UALR, Little Rock, Arkansas USA. UALR, Little Rock, Arkansas USA.

[6] 陈爽，宋金玉，刁兴春等. 基于马尔可夫模型的枚举型缺失值估计[J]. 上海交通大学学报. 2013, 47(8): 1246-1250.

[7] Ben-Gal Irad. Bayesian Networks[J]. Encyclopedia of Statistics in Quality & Reliability. 2008.

[8] 刘峰. 贝叶斯网络结构学习算法研究[D]. 北京邮电大学 博士学位论文, 2007.

[9] Callan R. Artificial Intelligence[M]. Palgrave Macmillan, 2003.

[10] Jie Cheng, David A. Bell, Weiru Liu. An algorithm for Bayesian belief network construction from data[C]. International Conference on Artificial Intelligence and Statistics, 1997: 83-90.

[11] Kuang Chen, Joseph M. Hellerstein, Tapan S. Parikh. Designing Adaptive Feedback for Improving Data Entry Accuracy[C]. ACM Symposium on User Interface Software and Technology(UIST), 2010: 239-248.

[12] Andre Inge. Hidden Markov Models: Theory and Simulation[D]. Bachelor Thesis, Stockholm University, 2013.

[13] Rabiner L., Juang B. H. An Introduction to Hidden Markov Models[J]. ASSP Magazine, IEEE. 1986, 3(1): 4-16.

[14] bnt-Bayes Net Toobox for Matlab[Z]. Available: code.google.com/p/bnt.

[15] Hidden Markov Model(HMM) Toobox for Matlab[Z]. Available: www.cs.ubc.ca/~murphyk/software/HMM/hmm.html.

[16] A. Hartemink. Banjo: Bayesian Network Inference with Java Objects[Z]. Available: www.cs.duke.edu/~amink/software/banjo/.