Using a Markov-Chain Model for Assessing Accuracy Degradation and Developing Data Maintenance Policies

Alisa Wechsler
Industrial Engineering and Management, Ben Gurion University, Beer Sheva, Israel., alisav@bgu.ac.il

Adir Even
Industrial Engineering and Management, Ben Gurion University, Beer Sheva, Israel., adireven@bgu.ac.il

Using a Markov-Chain Model for Assessing Accuracy Degradation and Developing Data Maintenance Policies

Alisa Wechsler
Dept. of Industrial Engineering and Management
Ben-Gurion University of the Negev
alisav@bgu.ac.il

Adir Even
Dept. of Industrial Engineering and Management
Ben-Gurion University of the Negev
adireven@bgu.ac.il

ABSTRACT

Accuracy reflects the extent of correctness of data. It is often evaluated by comparing the values recorded to a baseline perceived as correct. Even when data values are accurate at the time of recording – their accuracy may degrade over time, as certain properties of real-world entities may change, while the data values that reflect them are not being updated. This study uses the Markov-Chain model to develop an analytical framework that describes accuracy degradation over time – this by assessing the likelihood of certain data attributes to transition between states within a given time period. Evaluation of the framework with real-world data shows its potential contribution for key data-quality management tasks, such as the prediction of accuracy degradation, and the development of data auditing and maintenance policies.

Keywords

Data Quality, Accuracy, Currency, Markov-Chain Model.

INTRODUCTION

The quality of data and information resources has long been identified as a key issue in organizational information systems. The broad adoption of data driven decision-making culture in organizations has led to growing investments in data warehousing and business intelligence infrastructures, as well as to explosion in the volumes of data collected and managed. As data resource become a key asset - the attention to data quality issues is on the rise, and so are the concerts regarding the associated efforts and costs (Heinrich et al., 2009; Even et al., 2010).

Accuracy, the extent of data correctness, is among the most discussed data quality dimensions. A data item is considered to be inaccurate if its value doesn't match the correct real-world value (Even and Shankaranarayanan, 2007). Common causes for inaccuracies are errors in data acquisition (e.g., flawed data-entry) and processing (e.g., algorithmic calculation errors). A variety of solutions for reducing such errors have been proposed (Olson, 2003) – e.g., improving the design of data-entry screens, training end-users, and redesigning processes, such that chances of error are reduced. This study addresses another possible cause for inaccuracies – in many cases data is recorded and processes correctly; however, as the associated real-world entity changes over time, the data is not kept up-to-date. This issue is often referred to as currency or recency (Even and Shankaranarayanan, 2007; Heinrich et al., 2009) - a quality dimension that reflects failures to keep data items up-to-date. In this study, we observe accuracy and currency as related issues, as we address accuracies that are caused by failures to update data even when changes in the real-world entity require us to do so.

Identifying and correcting accuracy defects is a challenging and expensive task (Olson, 2003; Even and Shankaranarayanan, 2007). Telling that a certain item is inaccurate, only by observing its current value, is impossible in many cases. Usually some baseline for comparison is needed – either the real-world entity itself (e.g., surveying a person for validating his/her personal details), or another data source that was validated to by accurate (e.g., a dataset obtained from a reliable and up-to-date resource). Validated data sources are not always available, and auditing a large dataset against real-world entities may turn out to expensive. A few studies (e.g., Ballou and Pazer, 1995; Even and Shankaranarayanan, 2007; Gelman, 2010) have pointed out the tradeoffs between the desired goal of raising accuracy to the highest possible level versus the high costs involved, and proposed approaches for assessing these tradeoffs and setting the optimal accuracy-level target. However, even if some level of inaccuracy is acceptable – it is still critical to assess, monitor, and improve the level of accuracy, otherwise quality might degrade to a point that that data resource might become unfit to use.

This study contributes to that end by developing a Markov-Chain model that describes accuracy degradation over time. Markov-Chain models (Ross, 1996) have been used in a plethora of scientific and applicative contexts – e.g., Information science, Queuing theory, and Internet-page ranking; however, they have rarely been applied in the context of data quality management. We suggest that the developed model can contribute significantly to some important data quality management tasks – estimating accuracy when a baseline for comparison is unavailable or limited, predicting accuracy degradation of newly acquired data, and prioritizing accuracy auditing and improvement efforts. We next describe the model development and discuss its potential contribution, followed by an evaluation of the model with real-world data. Finally, we offer some concluding remarks and discuss limitations and directions for future research.

ACCURACY DEGRADATION MODEL

The model developed in this study, which aims at explaining and assessing accuracy degradation over time, addresses a tabular dataset with N records (indexed by [n]), each reflecting a specific instance of a certain database entity, and M columns (indexed by [m]), each reflecting a certain entity property. The model assumes that data values are accurate (i.e., reflects correctly the real-world attribute) at the time of recording, and remain unchanged else if updated later. Certain properties of real-world entity instances may change over time, and accuracy defects occur when a recorded data item is not being updated accordingly; hence, does not reflect correctly the current real-world value.

The Markov-Chain model of stochastic processes is based on the likelihood of a certain object to transition from one state to another within a given time period. We first address a single data item, which describes a certain property of a real-world entity. A data item may transition, within a given time period, from one state (a certain data value) to another. For example (Figure 1), the "Marital Status" of a certain person may transition between four states – "Single", "Married", "Divorcee", or "Widower". If a certain person is single (state 1) at the beginning of a certain time period, with some probability ($P_{1,1}$), s/he will stay single by the end of that period. However, with some probability s/he may become married, divorcee, or widower ($P_{1,2}$, $P_{1,3}$, and $P_{1,4}$, respectively). Similarly, we can define transition probabilities between all other states.

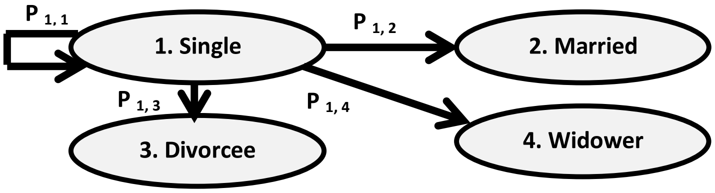


Table 1. Transition Probabilities for the "Single" state

The model targets data attributes with discrete value domains – i.e., a finite set of J possible values, indexed by [j] (e.g., Marital Status, Occupation, or Region of residence). We treat the time as a discrete variable (t = 0, 1, 2, ..), where the values reflects equal time intervals (e.g., day, month, year) . The fixed-size time intervals reflect, in our case, periodical data auditing - by the end of each time interval, we decide whether or not to audit and correct certain data values. Within a given time interval, the data value in attribute [m] may transition from state [i] to state [j] (or remain at state [i]) with a probability of $P^m_{i,\ j}$, such that $\sum_{j=1..J} P^m_{i,\ j} = 1$ for each [i]. The transition probabilities for attribute [m] can be represented in a form of a matrix P^m, and the model assumes that this matrix is identical for all the records in the dataset.

$$P^m = \begin{array}{cccc} {P^m}_{11} & \cdots & \cdots & {P^m}_{1J} \\ \vdots & {P^m}_{jj} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ {P^m}_{J1} & \cdots & \cdots & {P^m}_{JJ} \end{array} \qquad (1)$$

The set of matrices {Pm} may help assessing attribute volatility. An attribute [m] is said to be stable if all diagonal-cell values $\{P^m_{j,j}\}$ are nearly 1, while non-diagonal cells are nearly 0 (at the extreme case, the attribute is said to be stagnant – once its value is set, it stays permanent and it's accuracy won't degrade over time). An attribute [m] is said to be volatile when some diagonal cells $\{P^m_{j,j}\}$ are much smaller than one, while non-diagonal cells are substantially greater than 0.

The Markov-Chain model assumes that the transition matrix P^m is known, or can be reasonably estimated from data samples. We initially assume that P^m doesn't change over time; hence, P^m (t), the t-steps transition matrix of attribute [m] (i.e., the set of probabilities that a certain value in attribute [m] will change from state [i] to state [j] after t periods) is the t-power of matrix P^m: P^m (t) = $(P^m)^t$. The Markov-Chain model assumes “memory-less” transitions (Ross, 1996) – meaning that the probability of having a certain value $X^{n,m}{}_{t+1}$ in attribute [m] of record [n] at the end of period t+1, depends only on the transition matrix P^m, and on the value ($X^{n,m}{}_t$) at the end of period t, and not on earlier values

$$P^m\{X^{n,m}{}_{t+1} = j \mid X^{n,m}{}_t = i_t, \ldots, X^{n,m}{}_{0^m} = i_0\} = P^m{}_{ij}(X^{n,m}{}_{t+1} = j \mid X^{n,m}{}_t = i_t) \qquad (2)$$

A data item in record [n] and attribute [m] is said to be accurate if its value $X^{n,m}$ reflects correctly the real-world value. We assume that each data item was accurate when being recorded (t=0), and remains unchanged else if updated later. With no updates, $X^{n,m}$ is accurate at time t if the real-world value has not changed, or if changed and transitioned back to the original value. We define $A^{n,m}{}_j(t)$ as the expect accuracy of data item [n, m] at time t, given a current value of j. It equals to the likelihood that the real-world value is still j at time t:

$$A^{n,m}{}_j(t) = P^m{}_{jj}(t) \qquad (3)$$

Be averaging, we can assess the expected accuracy level of a record, an attribute, or the entire dataset ($A^{R(n)}(t)$, $A^{C(m)}(t)$ and A(t), respectively), given the set of known data values at time:

$$A^{R(n)}(t) = \frac{1}{M}\sum_{m=1}^{M} A^{n,m}{}_j(t), \qquad A^{C(m)}(t) = \frac{1}{N}\sum_{n=1}^{N} A^{n,m}{}_j(t), \qquad A(t) = \frac{1}{NM}\sum_{m=1}^{M}\sum_{n=1}^{N} A^{n,m}{}_j(t) \qquad (4)$$

The time for transitioning out of state j, can be approximated by an exponentially distribution, where $\lambda^{n,m}{}_j$ is the rate of data item [n, m] to leave state j and α is an auxiliary parameter (Ross, 1996).

$$F_j(t) = 1 - \exp\{-\alpha\lambda^{n,m}{}_j t\} \qquad (5)$$

Until time t, if the real-world value had not transitioned, the associated data item is accurate. Therefore, using the approximation, the expected dataset accuracy can be expressed as:

$$A(t) = \frac{1}{NM}\sum_{m=1}^{M}\sum_{n=1}^{N} \exp\{-\alpha\lambda^{n,m}{}_j t\} \qquad (6)$$

Using the exponential approximation for evaluation is substantially less time consuming than using the model, as it doesn’t require matrix multiplication; hence such an approximation has importance in cases where the dataset has a large number of records, or where the prediction involves a large number of time periods. It can be shown that the expected accuracy is a number between 0 and 1, where the averaging method adheres to the guidelines in (Even and Shankaranarayanan, 2007). With different relative importance of records and attributes, these definitions can be extended to use a weighted-average formulation proposed in that work.

The proposed model can support important data quality management tasks: ***(a) Estimating accuracy level*** – measuring the accuracy of a large dataset is challenging and expensive, as it requires a baseline for comparison (e.g., the real-world values, or another dataset that was validated for correctness). The suggested model permits accuracy estimation of a dataset, or subsets within, given the current values and the time since their last update, without requiring assessment against a baseline. ***(b) Predicting future accuracy degradation*** – when new items are recorded, the model can help predicting their accuracy behavior over time. Knowing the recorded value j, the model can help predict the accuracy level at time t, or the time until the accuracy will go below a certain desired threshold value, and ***(c) Prioritizing data maintenance efforts*** – the model may help assessing the accuracy behavior of data subsets (records and/or attributes), and setting auditing and maintenance priorities accordingly. As discussed earlier, the set of transition matrices can help differentiating between stable versus volatile attributes. Further, given the current data values, records with higher inaccuracy likelihood can be detected.

A key challenge with the proposed framework is estimating the transition matrices $\{P^m\}$. Such estimation requires a large-enough sample of data records, which includes the history of data-value transitions. Given such a sample, the estimation can be done in a relatively simple manner – the matrix component $P^m{}_{i,j}$ will be estimated by the number of times that attribute [m] transitioned from value [i] to value [j], divided by the total number of attribute [m] transitions from of value [i] (including “transitions” from [i] back to [i]). The next section demonstrates a case in which the availability of such a data sample permitted reasonable estimation of transition matrices and prediction of accuracy behavior over time.

EVALUATING THE MODEL WITH REAL-WORLD DATA

The evaluation described in this section used a dataset published by the Central Bureau of Statistics. The dataset contains economic-performance indicators on 124 industrial sectors, which were collected annually over a 14-year period (total of 3224 records). In accordance with our modeling assumptions – the data items were updated in fixed time intervals and, as the data source is considered highly reliable, it was reasonable to assume that the numbers provided are accurate.

We evaluated three financial attributes – Sales, Revenue, and Export. These characteristics have continuous value domain, and had to be "discretized". This was done by classifying each industry/year record 10 equally-size deciles, based on the value range of each indicators (e.g., the "Pharmaceutical" industry is in the 2nd decile in terms of Sales, 4th in Revenue, and 1st in Export), this division relies in the assumption of uniform distribution of the values. Over time, some industries improved their ranking, while others declined; hence, if one uses data records that were published a few years back –these records will not reflect the accurate ranking. The extent of inaccuracy degradation over a time period of Z years can be therefore assessed by comparing the ranking in records from year Y to records from year Y+Z, where a record is considered inaccurate when at least one of the three rankings has changes. Our analysis reflects two manifestations of the time variable: a) Learning: the number of periods between the first update and the last update, and b) Prediction: the number of periods between the year of last update and the year of accuracy assessment

To estimate the transition matrices, we used 80% of the records as a training set, and the rest 20% as a test set. We repeated this process 10 times with different random permutations and averaged the results. To assess performance, we used the commonly-used Kullback-Leibler Distance (KLD) metric (Do, 2003), where the lower is the KLD, the better is the prediction. Here, we use it to measure the distance between the predicted accuracy level APR(t) versus the actual AAC(t):

$$KLD(t) = DA^{AC}(t)\cdot\log\frac{DA^{AC}(t)}{DA^{PR}(t)} + \left(1-DA^{AC}(t)\right)\cdot\log\frac{\left(1-DA^{AC}(t)\right)}{DA^{PR}(t)} \qquad (7)$$

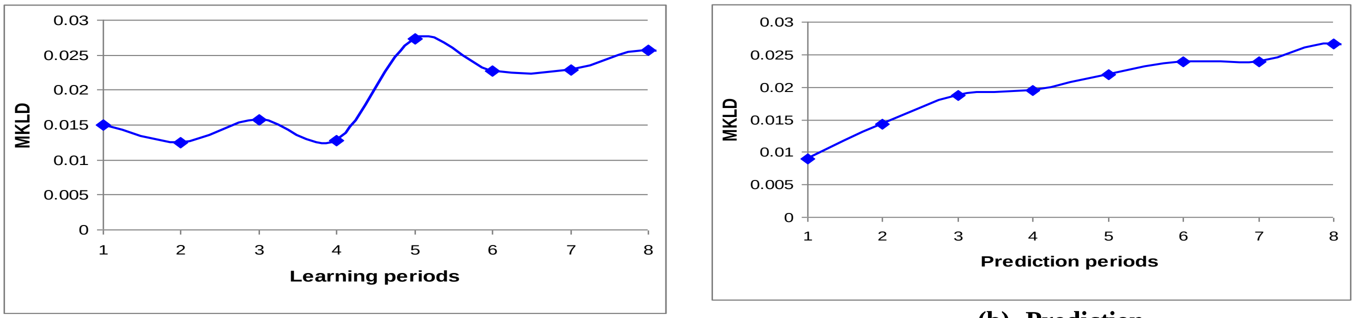
Where,

KLD(t): The Kullback-Leibler Distance at time t

AAC(t): The true dataset accuracy at time t, where last update is at t=0

APR(t): The predicted dataset accuracy (the model's output) at time t

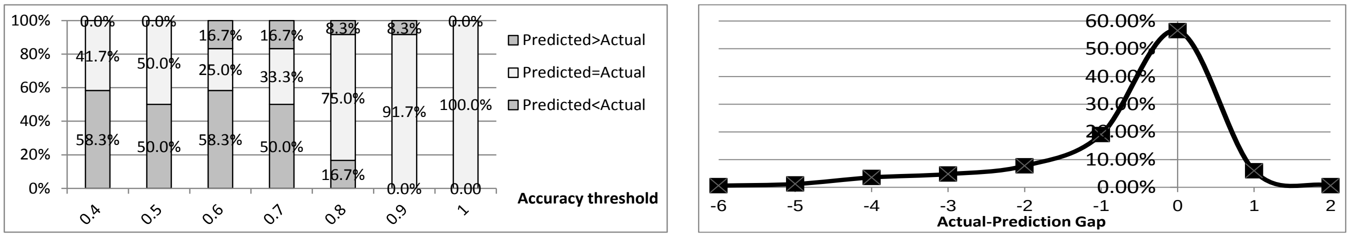
Following these analysis principles, we used the dataset to assess the potential contribution of the model for the data quality management tasks discussed earlier. To estimate current accuracy levels, the Mean KLD (MKLD) of the ten predictions was calculated versus learning (Figure 2a) and prediction (Figure 2b) times. The MKLD values are relatively small (less than 0.027), reflecting strong performance. As expected, both learning and prediction performance degraded with a larger number of periods. Prediction models usually perform better with short-term predictions. Further, in this particular case - transition matrices {P^{m}} were assumed permanent over time; however – it is reasonable to assume that over a long period of time, the transition behavior will change; hence, learning over a too-long period might hinder prediction capabilities.



(a) Learning **(b) Prediction**

Figure 2. Mean KLD (MKLD) versus Time

To assess the prediction of future accuracy degradation, we asked: given a certain threshold, can we predict the number of periods that it will take the accuracy of a perfectly-correct dataset to decline below that threshold. We assessed the predicted number versus the actual for thresholds ranging between 0.4 and 1. The results (Figure 3a) show that in the majority of cases the prediction was either identical to the actual or lower. The gap distribution (Figure 3b) shows that in 56% of the cases, overall, the prediction was precise, in 37% it was lower than the actual, and only in 7% the prediction was higher than that actual. These results have important data quality management implications, as the model is shown to be stringent – in the majority of cases we would have audited data items on time or earlier.



(c) Prediction versus Actual per Threshold **(d) Prediction-Actual Gap Distribution**

Figure 3. Accuracy Degradation – Prediction versus Actual

Finally, we examined the suitability of the exponential approximation of accuracy degradation. For each record, we calculated the average parameters that minimize the error between the model prediction and the exponential approximation for different prediction periods between 1 and 12. As can be seen (Figure 4), the exponential approximation of accuracy is only slightly below the model's prediction.

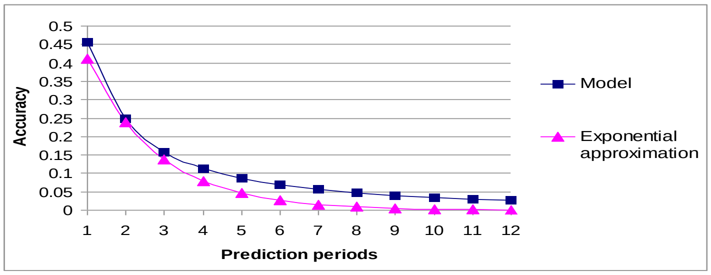


Figure 4. Exponential model and the Markov model predictions

CONCLUSIONS

This study describes the development of a model for simulating and predicting accuracy-degradation behavior over time, using the Markov-Chain modeling approach. As discussed and demonstrated in our evaluation with real-world data, the model has a high contribution potential within a few important data quality management tasks – estimating accuracy levels, predicting accuracy degradation, and prioritizing accuracy maintenance efforts. The results obtained in our preliminary evaluation confirmed the potential contribution of the model. Obviously, these are just the preliminary results and some additional evaluation is required with real-world datasets in other data management contexts. Notably, Markov-Chain models have rarely been applied in the context of data quality management, and we see it as a contribution of its own. As the study is still progressing, we intend to explore this direction further, and we see it as promising avenue for developing methodologies and tools for aiding data quality management efforts.

Obviously, some of the modeling assumptions made do not hold "as is" in many real-world cases. Data audits and corrections are not always done at fixed-length time intervals; what required a different modeling of the time variable. Data attributes (e.g., salary, length, and duration) are often continuous, and in some real-world scenarios, "discretization" into a set of bins, as done in this study, is not a valid solution. Further, the transition matrix Pm may not be identical for all records (e.g., when it depends on the value of other attributes), may change over time, and may not adhere to the "memory-less transitions" assumption. Notably, a plethora of Markov-Chain model studies have addressed such limitations and the solutions offered can be adopted for extending the framework developed here at later stages.

REFERENCES

1. Ballou, D. P., and Pazer, H. L. (1995). Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff. *Information Systems Research*, 6, 1, 51-72
2. Do, M. N. (2003) Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models, *IEEE Signal Processing Letters*, 10, 4, 115-118
3. Even, A., and Shankaranarayanan, G. (2007) Utility-Driven Assessment of Data Quality, *The DATA BASE for Advances in Information Systems*, 38, 2, 76-93
4. Even, A., Shankaranarayanan, G., and Berger, P.D. (2010) Evaluating a Model for Cost-Effective Data Quality Management in a Real-World CRM Setting, *Decision Support Systems*, 50, 1, 152-163
5. Gelman, I. A. (2010) Setting Priorities for Data Accuracy Improvements in Satisficing Decision-Making Scenarios: A Guiding Theory, *Decision Support Systems*, 48, 4, 507-520
6. Heinrich, B., Kaiser, M. and Klier, M. (2009) A Procedure To Develop Metrics For Currency And Its Application In CRM, *ACM Journal of Data and Information Quality*, 1, 1, 1-28
7. Olson, J. E. (2003), Data Quality: The Accuracy Dimension, Morgan Kaufmann Pub.
8. Ross, S. M. (1996) Stochastic Processes, Wiley.