The 3rd International Symposium on Frontiers in Ambient and Mobile Systems (FAMS)

A Histogram Method for Summarizing Multi-Dimensional Probabilistic Data

Ashraf Iqbal\textsuperscript{a}, Hai Wang\textsuperscript{b}, Qigang Gao\textsuperscript{a}

\textsuperscript{a}Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

\textsuperscript{b}Department of Finance, Information Systems, and Management Science, Saint Mary’s University, Halifax, NS, Canada

**Abstract**

Currently, many database applications deal with large imprecise and uncertain datasets. Probabilistic data summarization has recently emerged and has already become an active research area in the database community. In this paper, we propose a data summarization method to summarize multidimensional probabilistic data using histograms. The proposed method iteratively constructs a histogram to represent the probabilistic data while maintaining a trade-off between minimizing the relative entropy among probability distributions and minimizing the space used by the histogram. The experimental results show that the proposed method achieves small errors for various compression ratios.

© 2013 The Authors. Published by Elsevier B.V. Open access under CC BY-NC-ND license. Selection and peer-review under responsibility of Elhadi M. Shakshuki

**Keywords:** Probabilistic Database; Uncertain Database; Data Summarization; Histogram; Wavelet;

1. Introduction

Uncertainty and impreciseness is a very common phenomenon in many real-world data, and probabilistic database has recently drawn much attention with its offer of representing such uncertainty. Probabilistic databases have already been found useful in many applications \cite{4-5, 8-15, 17, 23} including in large data integration, in information extraction from noisy data in the web, in RFID and other sensor data management, in data cleaning for data warehouses and data mining tasks, in social networks, in data management, in scientific data management, in risk prediction etc. However, despite their benefits, probabilistic databases also bring forward some associated challenges such as higher scalability to represent all possible worlds for probability distribution, and difficulty in representing the semantics for underlying uncertainties in a relation.

Although there exists a number of data summarization techniques for standard (deterministic) databases, histograms are the most popular among them. In this paper, we propose a histogram method to
summarize multidimensional data represented as a tuple-independent probabilistic model. The proposed method, using a dynamic programming approach, iteratively splits histogram buckets until the given space limit is reached. The optimal splitting decision is based on a trade-off between minimizing the relative entropy among probability distributions, and minimizing the space occupied by buckets. Finally, the probabilities for each leaf buckets are estimated either using the Haar wavelet technique, or just as the mean of probability distributions in the bucket.

We conducted experiments for multi-dimensional real-world datasets, and measured the error rates using some benchmark error functions. We found that the proposed method yields small errors which converge for various compression ratios.

2. Background and Related Work

2.1. Probabilistic Database Models

There exist a few prototypes for probabilistic databases [1-3, 18-19, 21]. Suciu et al. [20] classified probabilistic database models into three broad categories based on their level of uncertainties: tuple-independent databases (where each tuple has an associated probability assigned, and is independent to any other tuples in the database), block-independent-disjoint (BID) databases (where tuples in a block are disjoint and individual blocks are independent of each other), and U-databases (where each possible combination of attribute values from the possible world is represented with an associated probability). Formally, a tuple-independent probabilistic database can be defined as a relation, \( R(\bigcup_{t_i \in R} t_i) \), where each tuple \( t_i \) is assigned to a probability, \( Pr(t_i) \), and can be represented using the \( m \) dimensions present at the relation, \( R \), as \( t_i = (d_{i1}, d_{i2}, \ldots, d_{im}) \).

2.2. Problem Statement

Our aim of this work is to construct a histogram for a given tuple-independent database schema, and a space constraint, while minimizing errors among the probability estimates of buckets and the actual probabilities of tuples in it. Formally, we can define the problem as follows:

Given a multi-dimensional probabilistic database relation, \( R(\bigcup_{t_i \in R} t_i) \) and a compression ratio, \( C_R \), we have to construct a histogram, \( H(\bigcup_{b_j \in R} b_j, Pr(b_j)) \) such that:

- \( \sum_{j=1}^{m} \text{Size}(b_j) \leq \text{Size}(R) \times C_R \) and,
- \( \sum_{j=1}^{m} \sum_{k=1}^{\text{Size}(b_j)} \text{Err} \_ \_ \text{func}(Pr(b_j)_k, Pr(t_{j,k})) \) is minimized.

Here, \( t_i \) is a tuple in relation \( R \) in the form \( d_{i1}, d_{i2}, \ldots, d_{in} \), where each \( d_{il} \) represents the value of \( l \)th dimension (i.e., attribute), and \( n \) is the total number of dimensions in the multi-dimensional probabilistic relation, \( R \). Each \( b_j \) represents the \( j \)th bucket, and \( m \) is the total number of buckets in the histogram. \( Pr(b_j) \) is the estimated probability value for all tuples \( t_{j,k} \) in the bucket \( b_j \), whereas \( Pr(t_{j,k}) \) is their actual probability. \( Err \_ \_ \text{func}(Pr(b_j)_k, Pr(t_{j,k})) \) is some form of error function, which calculates the differences between these two probability values.

2.3. Related Work

Although data summarization techniques have well been studied for standard (deterministic) databases, researchers have started dealing with this research problem for probabilistic databases only a couple years
back. The existing solutions are generally based on the particular probabilistic database models they are using based on the same techniques (e.g., histograms, wavelets etc.) people used for standard (deterministic) databases.

Cormode et al. [6] were one among the pioneers in addressing this research problem in the context of probabilistic database, and proposed a histogram-based approach where each bucket was represented using a probability density function (PDF). A dynamic programming approach was used to minimize some error metrics (e.g., sum-squared error, max error etc.) among these PDF estimates and the actual probability values. Cormode et al. [7] extended their work to better address the approximate queries while utilizing both histogram and Haar wavelet-based approaches. Optimal histogram buckets and wavelet coefficients, in their proposed method, were calculated using error metrics. Larusso et al. [16] also employed a PDF-based approach for the BID representation model of probabilistic database. They constructed such PDFs using Chebyshev and Minimax polynomial functions for a given range of error bounds. All these previously proposed techniques focused on one dimensional data.

3. The Proposed Data Summarization Method for Multi-Dimensional Probabilistic Data

We propose a histogram-based data summarization approach for multidimensional probabilistic database. We extended a popular histogram method [22] for tuple-independent probabilistic database model, while also making use of the Haar Wavelet technique for better probability approximation in each bucket. We are describing the components of the proposed method in next sub-sections.

3.1. Constructing the Histogram

We constructed a histogram for multidimensional tuple-independent probabilistic data using the following steps.

1. Initially consider the input data as a single bucket, and calculate the relative entropy for that bucket.
2. If the allocated space limit is reached or relative entropies for all buckets are zero, then, estimate probabilities for leaf buckets (as discussed in section 3.3). Go to step 3 otherwise.
3. Find the bucket with the maximum relative entropy. Find out the best splitting decision for that bucket using step 4.
4a) For each unique dimension value, consider splitting the bucket along that dimension value, and calculate relative entropies.
4b) For each cell containing a non-zero probability value, calculate relative entropy considering that cell as an outlier.
4c) Find out the optimal splitting decision using 4a) and 4b), for what we get the best trade-off between minimizing relative entropy, and minimizing the space occupied.
5. Split the buckets based on the decision found in 4c). Go to step 2.

As discussed in the above steps, at the initial step, the algorithm considers the whole probabilistic relation as a single bucket, and starts with the relative entropy calculated for that single bucket. After that, it iteratively finds the optimal splitting decision until either the space limit is reached or when all relative entropies are zero. At every such attempt, it tries to split the bucket with the maximum value of relative entropy. The splitting decision can be either for splitting the bucket into two new buckets, or to result into a singleton (i.e., individual tuples). Once the stopping criteria for iterations are met, the algorithm determines the estimated probability for each leaf bucket using an algorithm discussed in section 3.2.
3.2. Estimating Probabilities in Leaf Buckets

Once the histogram is constructed, we need to estimate the probability value for each leaf buckets. The proposed algorithm for estimating probabilities in leaf buckets is shown in Fig. 1.

| Algorithm: Estimate_Probability (R) |
|-----------------------------|
| Begin                       |
|   Prob_Est := 0.0;          |
|   Standard_Dev :=          |
|   Calculate the standard  |
|   deviation for the        |
|   probabilities in R;      |
|   If (Standard_Dev/Mean) <= \( \epsilon \) Then |
|     Prob_Est := Calculate  |
|     the estimated probability using Haar Wavelet; |
|   Else |
|     Prob_Est := Calculate  |
|     the mean for the        |
|     probabilities in R;    |
|   End If                   |
|   Return Prob_Est;         |
| End                         |

Fig. 1. The proposed algorithm for estimating probability in leaf buckets.

For a given portion of relation corresponding to a leaf bucket, the algorithm determines both mean and standard deviation of the probability values in that bucket. Haar wavelet technique generally provides better approximation for probability distribution with small standard deviations. Thus, in the proposed algorithm, we compared the ratio of the standard deviation and the mean against a threshold, \( \epsilon \) to decide whether Haar wavelet should be used to estimate the probability for the bucket.

4. Experimental Results

We conducted some experiments on real-world datasets and measured the error rates using some benchmark error functions for varying sizes of compression ratios. We could not find any existing method addressing the data summarization problem for multidimensional tuple-independent probabilistic data, and hence, we could not compare our results with others.

4.1. Datasets

We used two variations of the benchmark Census-Income (KDD) Data Set from the UCI repository (http://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29): Census-2D (containing attributes Age, and Family members under 18), and Census-3D (containing attributes Age, Family members under 18, and Occupation). In all experimental settings, we considered the cells as the input probabilistic relation, while calculating the probabilities as a ratio of their frequencies to the overall total frequency in all cells. This provides a practical probability distribution for the real-world to be abstracted.

4.2. Results

We employed a number of error measures to calculate the error rates of the estimated probabilities in histogram buckets, including the Sum-Squared-Relative-Error, Sum-Squared-Error, and Sum-Absolute-Error for tuple-independent probabilistic model.

We need to set some threshold values to conduct the experiments. For example, SSRE calculation depends on providing a suitable value for the smoothing variable, \( c \) [7]. We conducted all of our experiments considering the mean value of all probability distributions for \( c \), mainly because some
probability values in our cases are much smaller. We also considered the value of threshold, \( \varepsilon \) as 0.1 to determine when to use Haar Wavelet to estimate the probability values in a leaf bucket. The experimental results are shown in Fig. 2. All error measures reported in Fig. 2 are ratios of errors found for particular compression ratios by such errors found considering only single bucket for the whole probabilistic dataset.

As shown in Fig. 2, we found a similar behavior in both datasets for different error measures in the sense that they converge very quickly for a considerably small compression ratio. If we compare SAE and other error measures in the two datasets, we found that SSE values are greater, since they reflect the absolute differences between probabilities instead of squared differences. We found similar behaviors for SSE, and SSRE error measures for both datasets, in the sense that they both converged to a smaller error rate.

![Fig.2.Experimental Results](image)

Fig.2.Experimental Results (a) SSRE for the Census-2D dataset; (b) SAE for the Census-2D dataset; (c) SSE for the Census-2D dataset; (d) SSRE for the Census-3D dataset; (e) SAE for the Census-3D dataset; (f) SSE for the Census-3D dataset.

5. Conclusion

We have proposed a histogram method for summarizing multidimensional tuple-independent probabilistic data. While relative entropy is proved very powerful in many information theoretic and machine learning approaches, we found that this measure also provides very good results in histogram construction for probabilistic data. This is further supported by the experimental results where we
employed some other error measures, which converged to a small error rate for a considerably very small compression ratio. With the exponentially increasing amount of big uncertain data, we believe that such a data summarization method for probabilistic data is essential for query optimization, quick approximation in data mining and data warehousing applications.

References

[1] Aggarwal CC. Trio A System for Data Uncertainty and Lineage. Managing and Mining Uncertain Data 2009:1-35.
[2] Antova, L., Jansen, T., Koch, C., Olteanu, D. Fast and simple relational processing of uncertain data. IEEE 24th International Conference on Data Engineering 2008; 983-992.
[3] Boulos, J., Dalvi, N., Mandhani, B., Mathur, S., Re, C., Suciu, D. MYSTIQ: a system for finding more answers by using probabilities. Proceedings of the 2005 ACM SIGMOD international conference on Management of data 2005; 891-893.
[4] Burdick, D., Deshpande, P.M., Jayram, TS, Ramakrishnan, R., Vaitheyanathan, S. Efficient allocation algorithms for OLAP over imprecise data. Proceedings of the 32nd international conference on Very large data bases: VLDB Endowment; 2006: 391-402.
[5] Chung, P.T., Hsiao, H.H. Probabilistic Relational Database Applications for Biomedical Informatics. Advanced Information Networking and Applications-Workshops, 2008.AINAW 2008: 720-725.
[6] Cormode G, Deligiannakis A, Garofalakis M, McGregor A. Probabilistic histograms for probabilistic data. Proceedings of the VLDB Endowment 2009;2(1):526-537.
[7] Cormode G, Garofalakis M. Histograms and wavelets on probabilistic data. Knowledge and Data Engineering, IEEE Transactions on 2010;22(8):1142-1157.
[8] de Keijzer, A., van Keulen, M. IMPRECISE: Good-is-good-enough data integration. Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on: IEEE; 2008: 1548-1551.
[9] Dalvi N, Ré C, Suciu D. Probabilistic databases: diamonds in the dirt. Commun ACM 2009;52(7):86-94.
[10] Deshpande, A., Guestrin, C., Madden, S.R., Hellerstein, J.M., Hong, W. Model-driven data acquisition in sensor networks. Proceedings of the Thirtieth international conference on Very large data bases-Volume 30: VLDB Endowment; 2004: 588-599.
[11] Garofalakis MN, Brown KP, Franklin MJ, Hellerstein JM, Wang DZ, Michelakis E, et al. Probabilistic data management for pervasive computing: The data furnace project. IEEE Data Eng Bull 2006;29(1):57-63.
[12] Gupta, R., Sarawagi, S. Creating probabilistic databases from information extraction models. Proceedings of the international conference on Very large data bases-Volume 32, Number 2: VLDB Endowment; 2006: 965-976.
[13] Hassanazadeh O, Miller RJ. Creating probabilistic databases from duplicated data. The VLDB Journal—The International Journal on Very Large Data Bases 2009;18(5):1141-1166.
[14] Jayram T, Krishnamurthy R, Raghavan S, Vaitheyanathan S, Zhu H. Avatar information extraction system. IEEE Data Eng Bull 2006;29(1):40-48.
[15] Lafferty J, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001): 282-289.
[16] Larusso ND, Singh A. Synopses for probabilistic data over large domains. Proceedings of the 14th International Conference on Extending Database Technology: ACM; 2011: 389-400.
[17] Re, C., Suciu, D. Management of data with uncertainties. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management; 2007: 3-8.
[18] Sen P, Deshpande A, Getoor L. PrDB: managing and exploiting rich correlations in probabilistic databases. The VLDB Journal 2009;18(5):1065-1090.
[19] Singh, S., Mayfield, C., Mittal, S., Prabakar, S., Hambrusch, S., Shah, R. Orion 2.0: native support for uncertain data. Proceedings of the 2008 ACM SIGMOD international conference on Management of data: ACM 2008: 1239-1242.
[20] Suciu D, Olteanu D, Ré C, Koch C. Probabilistic Databases. Synthesis Lectures on Data Management 2011;3(2):30-40.
[21] Wang DZ, Michelakis E, Garofalakis M, Hellerstein JM. BayesStore: managing large, uncertain data repositories with probabilistic graphical models. Proceedings of the VLDB Endowment 2008;1(1):340-351.
[22] Wang H, Seveik KC. Histograms based on the minimum description length principle. The VLDB Journal 2008;17(3):419-442.
[23] Wenzhong Zhao, Probabilistic database and their applications. PhD thesis, University of Kentucky, 2004.