Mining Evolving Customer-Product Relationships in Multi-Dimensional Space[*]

Xiaolei Li Jiawei Han Xiaoxin Yin Dong Xin

University of Illinois at Urbana-Champaign, USA

Abstract

Previous work on mining transactional database has focused primarily on mining frequent itemsets, association rules, and sequential patterns. However, interesting relationships between customers and items, especially their evolution with time, have not been studied thoroughly. In this paper, we propose a Gaussian transformation-based regression model that captures time-variant relationships between customers and products. Moreover, since it is interesting to discover such relationships in a multi-dimensional space, an efficient method has been developed to compute multi-dimensional aggregates of such curves in a data cube environment. Our experimental results have demonstrated the promise of the approach.

1. Introduction

Previous studies such as association mining [2] or sequential pattern mining [3] on transaction databases have been focused primarily on *item-item relationships*. However, many businesses may like to find interesting *customer-product relationships*, especially their evolution with time. Such analysis, though related with time, has some fundamental differences from time-series analysis [1]: (1) transactions are often irregular and sparse, and (2) they carry different semantic meaning. In addition, it is highly desirable to analyze such time-variant transaction data in a multi-dimensional space in an OLAP manner [5].

To perform a systematic analysis at mining such relationships, we propose a Gaussian transformation-based regression analysis framework, ***Gaure***, for analyzing transaction databases. *Gaure* preserves the semantics of time-variant transactions, integrates transaction analysis and time-series analysis methods, and facilitates multi-dimensional trend analysis of transaction-series.

2. *Gaure*: Gaussian Transformation Based Regression

The input to the problem is a *transaction series* database, T, consisting of three dimensions: (1) customer, (2) item, and (3) a sequence of *transaction units* that correspond to transactions between customer and item. For each customer and item pair, (c_i, i_j), there is a sequence of transaction units: $\langle (t_0, m_0), (t_1, m_1), \ldots, (t_k, m_k) \rangle$, where m_l is its measure at time t_l. One could view each sequence as time series data and perform traditional regression to mine it [6]. However, there are several problems with this approach.

First, typical time series data have measures taken at regular intervals. In contrast, measures in transaction databases occur irregularly. Second, in traditional regression, the interested function is assumed to be stable in between samples. In transactions, the measure is semantically zero between transactions, and this zero value actually affects the mining. Third, transactions might be very sparse in time. The impoverished nature of the data does not lend well to typical regression. Fourth, small transactions that occur closely in time are equally interesting as one big transaction. To resolve these problems but yet still leverage the methods developed in statistics, we introduce a novel method called *Gaure*: Gaussian Transformation Based Regression. The method has two major steps: Gaussian transformation and regression.

Gaussian Transformation: The transformation proceeds as follows. For every (t_i, m_i) transaction unit, create a Gaussian distribution function with a mean of t_i and variance of σ^2. m_i will be used as a scalar multiplier on the function. Formally, for every (t_i, m_i), create a function of $f(t) = \frac{m_i}{\sigma\sqrt{2\pi}} e^{-(t-t_i)^2/(2\sigma^2)}$. Recall that each row in T contains a sequence of transaction units. For each such sequence, add the set of $f(t)$'s together, and we get the following function $g_{c_i,i_j}(t)$ for every (c_i, i_j) row in T. Equation (1) is the Gaussian Transformation Function (GTF).

$$g_{c_i,i_j}(t) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{i=0}^{k} m_i e^{-(t-t_i)^2/(2\sigma^2)} \tag{1}$$

[*] The work was supported in part by National Science Foundation under grants IIS-02-09199/IIS-03-08215. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Regression: The GTF transforms and smoothes the discrete transaction data into a numerical function. Unfortunately, it is cumbersome to keep all the parameters of Equation (1) for every row in T. To ameliorate this, we sample the GTF in its Equation (1) at various points and use them to find the least squares fitting polynomial. This new function will be known as the Gaussian Transformed Regression Function (GTRF).

The GTRF preserves the timings of the transactions and enhances the desired transaction semantics. The semantics are enhanced through the additive nature of GTF. When several transactions occur close in time, their individual Gaussian functions will sum to a bigger function. This effect is actually appropriate for the shopping pattern semantics. The GTF and GTRF will reveal these sorts of patterns while traditional regression based on the original points will miss them entirely. We show a concrete example to illustrate the argument. In Figure 1, the original sequence of shopping transactions with their measures are shown as points. In terms of shopping patterns, the small purchases at around 20 are just as interesting as the bigger ones. In the same figure, the curve shows the resultant GTF using $\sigma = 2$. As one can see, the small purchases added up to a hump bigger than the earlier ones.

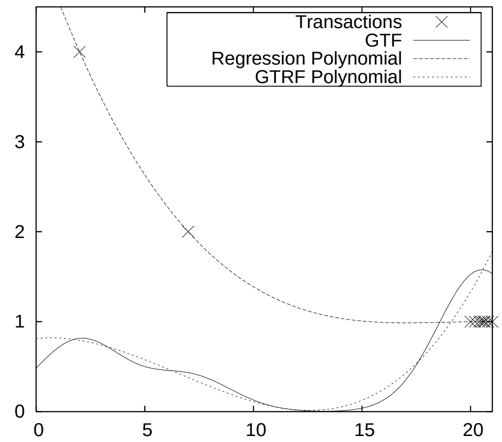

Figure 1. Transaction points, GTF, regression functions, and GTRF with 3rd order polynomial.

Figure 1 also shows 3rd order polynomial fitting functions of the transaction sequence. The "Regression Polynomial" shows a function that is fitted on the 10 data points alone. The "GTRF Polynomial" shows a 3rd order polynomial fitted using 20 evaluations of the GTF at $x = \{2, 3, \ldots, 20, 21\}$. The "Non-zero Polynomial" did its job of fitting the original data perfectly but missed the semantics. On the other hand, "GTRF Polynomial", did a very good job of representing the data and its semantics.

3. Multi-Dimensional Analysis

Multi-dimensional analysis can be performed efficiently on the Gaussian-based transformation and regression. This is accomplished by constructing a data cube using the coefficients of the GTRF polynomial. Recall that a data cube consists of all possible group-by's across all dimensions. Theorem 3.1 shows how to compute these aggregated group-by's. Furthermore, it implies for any cell in the data cube, its GTRF polynomial coefficients derived by the SUM operator from its descendent cells are exactly the same as the ones derived by a least squares solver if given the original data [4]. Thus, in the least squares sense, all cells in the data cube have the minimum error.

Theorem 3.1 (GTRF Aggregation) *Let there be m cells ($c_1, c_2, \ldots, c_m$) from the regression coefficient data cube with their respective k-th order GTRF coefficient vectors ($a_1, a_2, \ldots, a_m$). The k-th order GTRF of the m cells' aggregate cell has coefficient vector $a_A = \sum_{i=1}^{m} a_i$.*

4. Conclusions

In this paper, we have presented a new framework to model transaction data that uncovers temporal customer-item relationships in a multi-dimensional space. To properly mine these patterns, we introduced a novel approach *Gaure* that applies Gaussian transformation-based regression to model transaction data evolving with time. The semantics of the evolution of shopping transactions are enhanced. Furthermore, the method facilitates efficient data cube-like aggregations in multi-dimensional space.

References

[1] R. Agrawal, K.-I. Lin, H.S. Sawhney, and K. Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In *VLDB'95*.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *VLDB'94*.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In *ICDE'95*.

[4] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-dimensional regression analysis of time-series data streams. In *VLDB'02*.

[5] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Mining and Knowledge Discovery*, 1:29–54, 1997.

[6] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In *KDD'01*.