Integrating and Mining Distributed Customer Databases

Ira J. Haimowitz Özden Gür-Ali Henry Schwarz

General Electric Corporate Research and Development

One Research Circle, Niskayuna, NY 12309

E-mail: {haimowitz, gurali, schwarz} @crd.ge.com

Abstract[1]

Large corporations often have different subunits sharing common customers, yielding distributed customer databases. Corporate risk and marketing functions seek areas where there is unusually high risk, or where one can target market. We present a three-phase process to solve this problem. First, we merge the distributed databases using decision tree induction into a database of unique customers, labeled by location, industry code, and financial parameters. Second, we reduce the customer table to three explanatory business factors and various outcome measures. An ANOVA Model identifies outstanding effects and outliers. By incorporating both main and interaction effects, this approach identifies outliers that are more likely to be "interesting" than would be found using only main effects. Third, we display the aberrations as peaks or valleys so a user can isolate opportunities. This framework approximates an "interestingness filter."

KEYWORDS: outliers, interestingness, ANOVA model, record matching, decision trees, business.

Motivations

Mining Distributed Customer Data for Outliers

Large corporations often have different subunits with common customers, but in distributed databases. Their customers cover many industries and geographic regions. Corporate risk and marketing functions wish to learn of customer types where there is unusually high risk, or where one can cross-sell related products. Examples of commercial customers include:

1. Business customers leasing computers, furniture, and copiers from subunits of the same equipment lessor.
2. Business customers enrolled in different types of phone service from subsidiaries of a telecommunications firm.
3. Business customers using different financial instruments from subunits of the same bank.

A company can pool and mine the customer databases of their subsidiaries to answer these questions:

- Which locations, industries, or combinations thereof, generate unusually high or low revenue?
- Which product lines of the corporations are most active in certain locations and industries?
- Which locations and industries are demonstrating significant recent activity?

Outlier Detection and Interestingness Criterion

Aberration mining, or outlier detection, is a well-studied area in statistics. In the univariate case the common practice is to declare points outside the mean +/- k standard deviations to be outliers. However, more robust analyses based on order statistics are available [Hoaglin83]. Outliers in multivariate settings can be similarly identified by considering the Mahalanobis distance from the origin [Johnson90]. A more general alternative is to cluster the data and declare the points that are far away from any cluster to be outliers.

In the Knowledge Discovery area, Matheus and Piatetsky-Shapiro's KEFIR system [Matheus95] applies a sequence of filters to find interesting knowledge in corporate health claims data. The intermediate "aberration filter" is implemented by examining significant differences from previous values in time. The final "interestingness filter" is based on knowledge being *actionable*, based in part on domain knowledge.

Our approach determines outliers by combining shallow domain knowledge of important business factors with statistical modeling of historical customer data. Deviations from the model describe differences from history.

Three Phase Architecture

Figure 1 illustrates our three phase architecture.The validate and match phase takes as input distinct customer data sets for different products that are in a standard file format. Our work in this area is similar in spirit to the record merging work of Hernandez and Stolfo [Hernandez95]. This phase merges records and creates a *match results data warehouse* with four tables:

1. *Customers* includes name, address, and total business activity, and is mined in the second phase.
2. *Deals* has transactions associated with each customer
3. *Pending records* holds customers difficult to match.
4. *Bad records holds* input data that is either empty or inconsistent (e.g city vs. state conflict).

A user may examine each table of the customer match results database using an OLAP-style *query tool,* with specialized screens for each table.

The summarize and mine phase accesses the customer table and summarizes the records into a table where each row specifies a unique combination of location, industry, and business subunit. Each row also contains aggregate

measurements such as total sales. This summary table is mined using an ANOVA model to produce outliers in both one and two dimensions.

The display phase takes the model results and visualizes the outliers graphically, with relative magnitudes.

In our implementation, the validate and match phase is implemented in Visual Basic and C, and creates a Sybase database. The summarize and mine phase is implemented in Sybase and in SAS. The display phase is implemented in Cornerstone. Each phase is automated, but the overall process is not. The architecture is modular, allowing different tools for database management, statistics, or display.

Merging the Customer Databases

The validate and match phase (U.S. patent pending, [US96[) begins with an empty customer table, and merges each subunit customer table in turn, removing duplicates. This phase reads each record, and validates (checks for quality) and normalizes each field to a standard form. The matching process produces lists of great and mediocre matches, and decides to either create a new customer record, update an existing record, or save the data in the pending table for human resolution.

Attribute by Attribute Matching Functions

Once the customer record has been normalized and checked for quality, it must be matched against entries in the Customer table. Due to the large databases, hash keys are used to identify a small set of customer records (the *candidate set)* against which the new data record will be matched.

Our principle to determine equal customers is whether the name and geographic locations match. The example below illustrates that two records for the same customer can appear as quite different strings:

IMAGING TECHNOLOGY INC
55 MIDDLESEX TURNPIKE BEDFORD MA
Zip: 017300000 Phone: 6179388444

IMAGING TECHNOLOGY
55 MIDDLESEX TPKE BEDFORD MA
Zip: 017301403 Phone: 6172752700

We developed matching algorithms for each attribute of a business record. Each algorithm computes a matching score from 0 (quite different) to 1 (the same). There were three classes of matchers:

1. Exact string comparisons (used rarely, e.g. country code).
2. Soundex-based matching for attributes like Name and City. Soundex is an algorithm used by telecommunications companies for matching phonetically similar names (e.g. Johnson, Johansen, Jonson, etc.).

Address matching supplements Soundex matching with up-front parsing and fitting to templates such as:

PO BOX Number Street name Street descriptor Direction

3. Weighted character comparison matching, for numeric attributes like Zip code and phone number

Machine Learning Approach to Matching

We hand-matched over 1100 pairs of close matching customer records with the attribute matching scores. A matrix of customer pair matches is illustrated below, where the last column designates whether the pair of records was a match (1) or a non- match (0):

name	addr	city	state	zip	cou	phon	cid	duns	uldu	ulna	sic	match
0.20	0.28	1.00	1.00	0.71	1.0	0	0	0	0	0.5	0.0	0
0.63	0.86	1.00	1.00	1.0	1.0	0	0	0	0	0.5	0.1	1

CART (Classification and Regression Trees) was used to obtain output of the form below, where a line with an asterisk indicates a rule. The first item in each row is the inequality, followed by the number of training cases, a distance measure, and a frequency of match. For example, line (4) below indicates that when name < 0.7 and phone < 0.905, there were 810 cases, and 0 percent of the training records matched.

1)root 1142 223.100 0.266200
2) name<0.7 818 2.989 0.003667
4) phone<0.905 810 0.000 0.000000 *
5) phone>0.905 8 1.875 0.375000 *
3) name>0.7 324 21.370 0.929000
6) addr<0.625 23 0.000 0.000000 *
7) addr>0.625 301 0.000 1.000000 *

We generated matching rules from frequencies close to

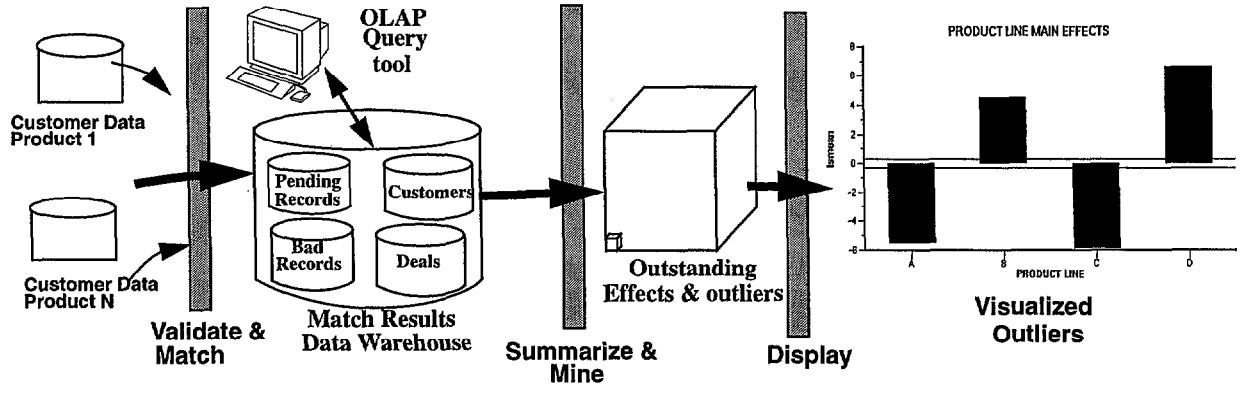

Figure 1 Three phase process for mining customer databases

0 or 1. Mid-range frequencies were indicative of needing a person's consultation. These rules have been supplemented with others suggested by users.

After matching, the customer table includes: location (e.g. Zip code range), 4-digit Standard Industry Code (SIC), products (i.e. business subunits), and total transaction amounts, (e.g. outstanding balance).

Summarize and Mine

The customer table is summarized by these business activity indicators for each industry, product line and location: number of sales contracts, average, standard deviation and sum of sales, average and standard deviation of length of contracts. This macro information is used to highlight unusual activity at different levels:

1. In a particular product line, industry or location.
2. For a particular product line and industry, or product line and location, or industry and location pair, adjusting for overall effects of product line, i
3. ndustry and location.
4. Too high or too little activity of a product line in a particular industry and location after considering main effects.

We formalize our process of calculating expectations and identifying outliers by modeling the business activity in our product-industry-location table with a multiplicative ANOVA type model:

$$\log_{10}\left(y_{ijk}\right) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \zeta_{ik} + \varepsilon_{ijk}$$

$i=1..n_{industry}, j=1..n_{product}, k=1..n_{location}$

y_{ijk}:business activity indicator for the cell product line j dealing in industry i and location k,

μ: mean business activity indicator y;

main effects beyond the mean α_i: general effect of industry i, β_j: general effect of product line j, γ_k: general effect of location k;

interaction effects beyond mean and main effects δ_{ij}: effect of product line j dealing in industry i, ζ_{jk}: effect of product line j dealing in location k;

residuals ε_{ijk}: residual of business activity indicator y_{ijk} after accounting for the main and interaction effects.

Especially for business activity indicators that can only be positive, multiplicative models have the advantage of not predicting negative values. The interaction effects to be included in the model should be chosen based on both interpretability, and significance of the term in the model. In our application the model fitting process resulted in a multiplicative model with main effects, and also interaction effects of product line with industry and with location. The model is fit with SAS Proc GLM.

The advantage of an ANOVA type model is that the dimensions of business activity need not be ordinal. For example it is very difficult to establish an ordinal scale for industry without resorting to auxiliary variables like total industry size. Another advantage is that ANOVA type models do not impose a linearity assumption on ordinal variables. For example, sales does not have to increase linearly with size of the industry.

Important in model construction is that there is valuable information in the fact that some cells of the cube do not have any sales contracts. If we are looking for a measure of total business activity and sales then we should represent it explicitly and generate observations with zero deals and zero sales. If the aim is to find the average size of a sales contract given that there is an activity, observations with zero average sales will be misleading.

Three levels of interestingness

Main effects and external information

The significant main effects help identify patterns like "Sales in the finance industry is significantly higher than in other industries." Patterns like "After taking into account the overall effects of industry, product line and location, product line C has little business in Mining" are discovered by comparing the predicted value for product line C sales for Mining based on the main effects multiplicative model with the fitted line C / Mining activity based on the interaction model. The difference is significant based on how many standard deviations (of the estimates) it is from zero.

We use the residuals to make statements like "after taking into account the general effects of product line, industry, location; and the product line dealing in a particular industry and location, it is unexpected that there are no sales contracts in zip code 51 in Retail by product line C." As a numerical example, assume that the multiplicative model has fit the following parameters:

$10^{\mu} = 50$, $10^{\alpha_{retail}} = 0.3$, $10^{\beta_{product\ line\ C}} = 2.1$,

$10^{\gamma_{zip\ code\ 51}} = 0.6$, $10^{\delta_{product\ line\ C,\ retail}} = 0.8$,

$10^{\zeta_{product\ line\ C,\ zip\ code\ 51}} = 1.2$,

$y_{retail,\ product\ line\ C,\ zip\ code\ 51} = 70$.

The observed business activity in product line C, location zip code 51 and industry retail is 70, whereas the expected one according to the model is: 50*0.3*2.1*0.6*0.8*1.2=18.1; quite different from 70. The significance of 70 is determined by it being outside 3 standard deviations in the distribution of residuals.

We present interesting main effects, and interaction effects and residuals, in a project with four product lines, 99 zip code locations and 10 industry groups. Figure 2 shows marked main effects of sales for industry and location; product line effects are in Figure 1, right.

The product line turned out to be the most influential factor in determining the sales of a product line in a location and industry. The horizontal reference lines are set at three times the std of the estimate of the effect. Note wholesale is within the lines, indicating that wholesale is not an outstanding industry. Also (Fig. 1), product lines D and B

have significantly higher sales compared to A and C. Location zip code 09 in NY State is comparable in size to the product line effects and indicates there is a severe lack of business activity in that location. Locations with significantly higher or lower sales are those with bars exceeding the lines. In terms of industries, Services and Manufacturing stand out as high sales sectors followed by Transportation and Retail. Agriculture, Mining and Public Administration have significantly low sales compared with other industries.

Interaction effects and main effect expectations

The second level of interesting information is obtained by assuming that management is familiar with the general trends as seen in the main effects. Significant interaction effects indicate pockets of unusual activity, defined by product line in industry or product line in location. Figure 3 illustrates results in this case. For example, although sales to the public administration sector are significantly lower than to other industries (as identified by significant main effects), there is a high degree of variability across product lines. The product line B has much lower sales than the others. Also, line B has an unusually high percent of its sales in zip code 06 and 00, compared with other product lines. These nuggets may be due to the nature of the product line or may indicate some best practices / improvement opportunities.

Residuals and expectations from all effects

Large residuals also have important information; they indicate that a product line in a particular location and industry has significantly higher or lower percent of sales than is expected based on the typical sales for that product line in that industry and in that location. In our project there are very few residuals beyond the three standard deviations limits; those outliers may be listed for a business manager.

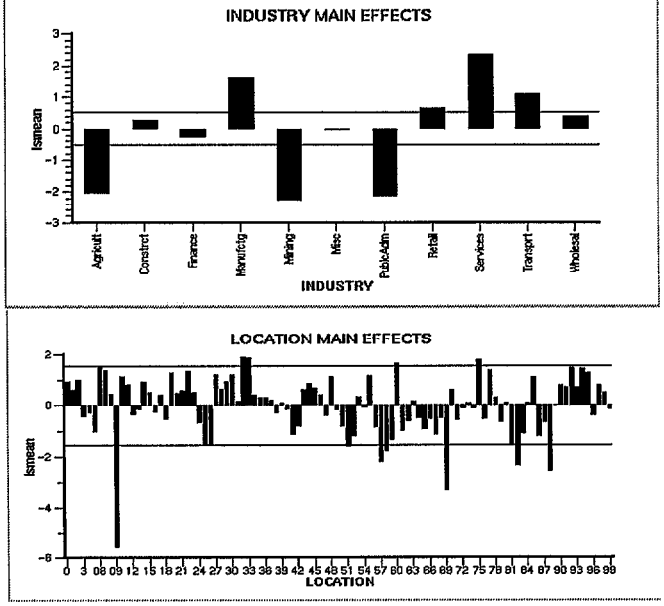

Figure 2 Significant main effects for sales.

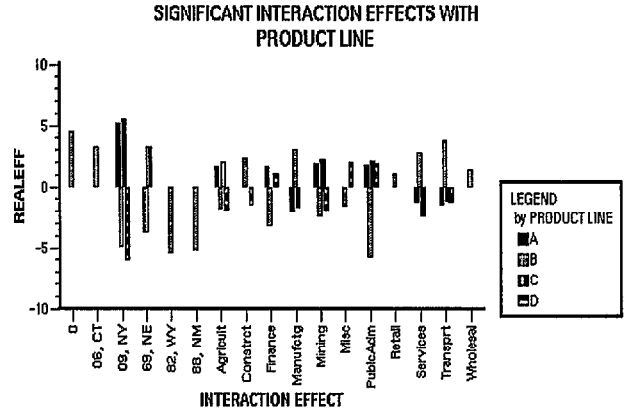

Figure 3 Significant interaction: product with location & industry.

Conclusions

Our approach is innovative in combining intelligent record matching followed by ANOVA modeling for detecting interesting aberrations. Business managers may know most main and some secondary effects; the issue is actionability. We can suggest actions by comparing significantly high and low industries and locations to corresponding current and leading economic indicators. Rules follow:

effect is significantly high			**effect is significantly low**		
leading indicators	current indicators		leading indicators	current indicators	
	high	low		high	low
high	no action	*current risk evaluation*	high	**marketing opportunity**	*upcoming marketing opportunity*
low	*risk alert: future*	**risk alert**	low	*may require quick action*	no action required

References

[Agrawal95] Aggrawal, R. and Psaila, G., "Active Data Mining," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, AAAI Press, 1995, pp. 3-8.

[Fayyad95] Fayyad U. M., Piatetsky-Shapiro G. and Smyth P., "From Data Mining to Knowledge Discovery: An Overview", in *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1995.

[Hernandez95] Hernandez, M. and Stolfo, S. "The Merge/Purge Problem for Large Databases." *1995 ACM/SIGMOD Conference.*, May 1995.

[Hoaglin83] Hoaglin D.C., Mosteller F., and Tukey J. W., *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, 1983.

[Johnson92] Johnson, R. A., *Applied Multivariate Statistical Analysis*, Prentice Hall: NJ, 1992.

[Matheus96] Matheus, C.J., and Piatetsky-Shapiro, G. "Detecting and Reporting What is Interesting: The Kefir Application to Healthcare Data," in *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1995.

[Myers95] Myers R. H., and Montgomery D. C., *Response Surface Methodology*, John Wiley & Sons, 1995.

[Searle71] Searle S. R., *Linear Models*, John Wiley & Sons, 1971.

[US96] Haimowitz, I.J., Murren, B.T., Lander, H., Pierce, B.A., Phillips, M.C., US patent application. 08/702379, "Method and System For Matching New Customer Records To Existing Customer Records In A Large Business Database," filed U.S. Patent/Trademark Office 8/23/96.