Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2003 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2003

Divide and Conquer: A Novel Approach to Segmentation and Model Building Using a Pattern-Oriented Clustering Approach

Yinghui Yang
University of Pennsylvania

Divide and Conquer: Novel Approach to Segmentation and Model Building Using a Pattern-Oriented Clustering Approach

Yinghui Yang
The Wharton School
University of Pennsylvania
yiyang@wharton.upenn.edu

Keywords: Data mining, customer profiling, signature discovery, clustering, local model

Introduction

Businesses gather enormous amounts of data in their day-to-day operations. Every interaction with a customer generates data and the amount of data gathered is rising exponentially. As technology advances, firms are able to track the origin of the transactions through customer identifiers such as cookies, shopper cards, credit card numbers, cell phone numbers etc. More and more, firms are realizing the importance of understanding and leveraging customer level data, and critical business decision models are built upon analyzing such data. For example, Amazon.com offers distinct home pages and recommends new products for customers based on personalization models built from data. Most credit card and cell phone fraud alerts are also issued from the analysis of customer level data. Consumer brand choice models and pricing models are heavily used in marketing endeavors.

While the expectation for customer level data analysis is high, there are still problems with existing analysis methods. Consumers still receive significant amount of mail advertising products that they are not interested in, and online recommendations are still far from perfect. In order to create more successful personalized systems and build more accurate consumer behavior models for customers, firms need to understand their customers better. This includes understanding customers' preferences through facts and customers' behavior through analyzing their transaction data. There has been much research done in this direction. For example, in the marketing literature, market segmentation approaches have often been used to divide customers into groups in order to implement different strategies. It has been long established that customers demonstrate heterogeneity in their product preferences and buying behaviors (Allenby and Rossi, 1999), and that the model built on the market in aggregate is less efficient than models built for individual segments. Our research approaches building more accurate customer models from the same perspective as market segmentation research. Our goal is to find the heterogeneity within customers' transactions, divide them into clusters and build more accurate models for individual clusters. Vast amount of market segmentation research focuses on examining how variables like demographics, socioeconomic status, personality and attitudes can be used in predicting differences in consumption and brand loyalty. But traditional market segmentation models often oversimplify the difference between segments by using changing model parameters to represent the difference. A drawback to this approach is that the criteria for segmentation is often simple, and needs to be specified upfront (e.g. segmenting based on "income") – a "hypothesis-driven approach for segmentation". An obvious extension is to use data-based clustering techniques for segmentation. While this practice is increasingly seen in segmentation, the actual clustering approaches that have been used (such as the k-means clustering technique) have been ad-hoc. It is generally not clear why a distance-based clustering in an *n* dimensional space, while convenient, is the appropriate manner to group customers. In our research we study a new approach to segmenting customers, one that is based on the idea that there may exist natural behavioral patterns in different groups of customers. For example, a set of behavioral patterns that distinguish a group of wireless subscribers may be:

- Their average call duration during weekday mornings is short and these calls are within the same geographical area.
- They call from outside the home area on weekdays and from the home area on weekends.
- They have several "data" calls on weekdays.

The above set of three patterns may be representative of a group of ‘consultants’ who travel frequently and exhibit a set of common behavioral patterns. Note that traditional hypothesis-driven segmentation and distance-based clustering are limited in their ability to learn this set as the distinguishing set of patterns for a cluster. If customers can be grouped into segments based on such natural behavioral patterns, it stands to reason that this may be superior to hypothesis-driven approaches and ad-hoc clustering approaches. This research proposes new methods for doing so, and studies whether grouping customers based on such patterns dominates conventional approaches.

Customers’ behavioral patterns can have different representations. A behavioral pattern can look like an IF-THEN rule. It represents what a customer will do under a certain circumstance. For example, if it’s during the weekend, customer *X* will make a call longer than one hour to California using his cell phone. We can also use a collection of things (called an “itemset” in the data mining literature (Agrawal et al. 1995)) a customer does together to represent certain behavioral patterns. For example, such patterns can look like the following. Customer *Y* buys cucumbers, tomatoes and tofu together. It’s also common to use sequences for pattern representation. One such example is that customer *W* visits cnn.com after msn.com. Another possible representation is a simple atomic condition based on variables. For example, the maximum amount of money Customer *Z* ever uses his visa credit card for a single transaction is $100.

Consumer transactions have natural categories and how to use patterns to discover those natural categories is a challenge. The reason is that many of these natural categorizations are not observable from the data. For example, web transactions may be for work, for entertainment, shopping for self, shopping as gifts, transactions made while in a happy mood, transactions made while in a not-so-happy mood etc. Customers do not indicate whether they’re happy or sad before starting a transaction. However, ignoring that they exist can result in learning incorrect or incomplete patterns. For example, when customer *X* is “relaxed”, (90% of the time), he often (80% of the time) buys healthy food such as salad, wheat bread etc. When he is “stressed” (10% of the time), he often (70% of the time) buys beer and chips. If we do not distinguish these two categories, we may only draw the following conclusion: Customer *X* buys healthy food 72% of the time (90% x 80% = 72%). The junk food purchasing behavioral pattern is usually overlooked, because the significance level is only 7% (10% x 70% = 7%). But this behavior is a very significant pattern that can help understand the consumer better. Since most of the rule discovery methods use a significance factor (called “support” in data mining literature(Agrawal et al. 1995)) to decide which rule to select, this problem of overlooking patterns can be pervasive.

In this dissertation, we propose a framework of using behavioral patterns discovered from customers’ transaction data to discover natural categories of transactions and further build more accurate customer behavior models. To do this, in this dissertation we propose a pattern-oriented clustering approach – that transactions can be clustered such that patterns generated from one cluster are significantly different from those generated from another cluster. We believe that each natural category of customer transactions can be characterized by its own distinguishing patterns. After we elicit multiple categories of customer transactions, we build one signature capturing the salient behavioral patterns for each category, and also one predictive model for each category. In the prediction stage, a new transaction is compared with all the signatures, and the closest signature is chosen. Then this new transaction is assigned to the category of transactions this signature represents and the model associated with this signature is used to predict for this transaction.

Our approach of incorporating behavioral patterns in modeling and prediction can be utilized in numerous real applications, since the phenomenon of behavior changes is pervasive. Building more accurate recommendation engines and detecting different types of fraud are certainly two natural application fields. What’s more are applications involving heterogeneous data records, such as shopping basket data, stock data, web browsing data etc.

The General Framework

This section describes the general framework (Figure 1) of this dissertation. The primary goal of this dissertation is to use behavioral patterns to build more accurate predictive models for customers. Our approach begins with a set of transactions. These transactions can be represented in different data formats as they are recorded, but they will be preprocessed so as to fit in the scheme of the behavioral pattern representation whichever is chosen. Then they are clustered into categories according to the inherent behavioral difference among the transactions. A signature will be elicited from each category of transactions after clustering. A local predictive model of each category is then built. The multiple local models are then used for prediction.

We divide this framework into three stages: clustering stage, signature discovery and model building stage, and prediction stage.

(a) *Clustering stage:*

The goal is to discover the natural categories among the transactions. Since there is no indicator showing which transaction belongs to which category and what exactly are those natural categories is unknown, this step is unsupervised learning. So, we need to have a reasonable method which can work towards getting relatively "good" clusters. This task itself is an important sub-problem within our framework. And we'll further illustrate this.

Before clustering, the representation for the behavioral patterns in the transactions needs to be chosen first. A representation has to be chosen to reflect the nature of the transactions. For example, itemsets are recognized as a good representation for customers' shopping behavior, since the shopping baskets contain items, and the co-occurrence of items reflects the customer's shopping patterns. This step is not automated. It involves domain knowledge from the application field where the data comes from and the understanding of the alternative pattern representations. The possible representations we consider include a set of variables, items, itemsets, sequences and rules, but not limited to these. The format of the signature and the model highly depend on the chosen representation of the behavioral patterns.

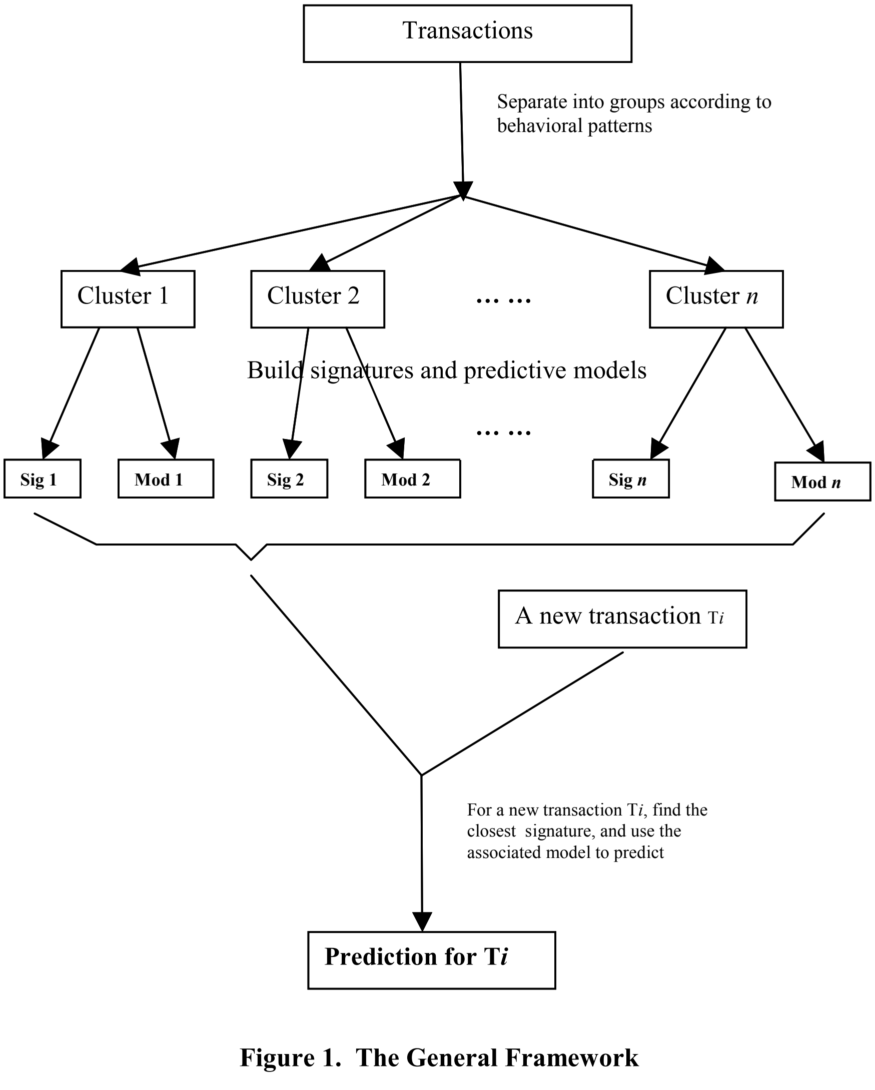
Transactions
Separate into groups according to behavioral patterns
Cluster 1
Cluster 2
… …
Cluster *n*
Build signatures and predictive models
… …
Sig 1
Mod 1
Sig 2
Mod 2
Sig *n*
Mod *n*
A new transaction T*i*
For a new transaction T*i*, find the closest signature, and use the associated model to predict
Prediction for T*i*

Figure 1. The General Framework

(b) *Signature discovery and model building stage:*

A signature should capture the salient behavioral patterns for each category. For example, a signature can be a set of significant itemsets in a category of transactions.

Besides describing the data with signatures, we build a local predictive model for each cluster. The dependent variable for the local model is what firms are interested in predicting, such as whether the customer will purchase a certain product or whether a certain transaction is fraudulent. And this part of the problem is supervised learning. The format of the model does not need to be the same as that of the signature. It can be any prediction model which fits the data well. Since signatures take the form which reflects the nature of the data, the format of the model also needs to be well chosen to reflect the nature of the data. For example, it's not appropriate to use a linear regression model for shopping transactions. Our choices for the local predictive models can include regression models, decision trees, neural networks etc.

(c) *Prediction stage:*

In this stage, we need to decide which model to use to predict for a new transaction. The signatures play a very important role in this selection procedure. Because a signature describes the characteristics of a category of transactions, comparing the new transaction with the signatures will give us a good picture about whether or not a new transaction comes from a certain category. After comparing the new transaction with all the signatures, we pick out the signature which is closest to the new transaction, and use the local model associated with that signature to predict for this new transaction. This requires the definition of closeness between a new transaction and a signature. Again, this definition may vary according to the representation of the signature. Taking itemsets as an example, the signature can be a set of frequent itemsets in a cluster, and the distance between a transaction and a signature can be defined as the number of itemsets which exist in both the new transaction and the signature.

Pattern-Oriented Clustering

The most essential part of our approach is to identify the natural categories of customer transactions. This is done via pattern-oriented clustering.. The goal for pattern-oriented clustering is to cluster transactions such that patterns generated from one cluster are "similar" to each other but "different" from patterns generated from other clusters.

Consider a collection of transactions to be clustered $\{ T_1, T_2, \dots, T_n \}$. Each transaction T_i contains a subset of a list of candidate items $\{ i_1, i_2, \dots, i_m \}$. A clustering C is a partition $\{ C_1, C_2, \dots, C_k \}$ of $\{ T_1, T_2, \dots, T_n \}$. Each C_i is called a cluster. The goal of our method is to maximize the difference between clusters and the similarity of transactions within clusters.

$$Maximize : M(C_1, C_2, ..., C_k) = Difference(C_1, C_2, ..., C_k) + \sum_{i=1}^{k} Similarity(C_i)$$

Both the difference and the similarity are defined in terms of the pattern representation. The chosen behavior representation affects the corresponding clustering method. We have developed a pattern-oriented clustering method based on itemsets. Because of space limit, we do not elaborate here. (For details, please refer to Yang and Padmanabhan (2003))

Related Work

At a high level, our approach advocates a "divide and conquer" approach to building models of consumer behavior since we first divide the data into groups and build separate models within each group. Even though no existing research has approached customer level data analysis from what's proposed in this dissertation, this spirit of "divide and conquer" is not new and there are several such ideas in the modeling literature.

As we mentioned earlier, huge amount of such "divide and conquer" work has been conducted in market segmentation community (for a review, please see Allenby and Rossi (1999)). And it has been established that this tactic can help build much better customer models. Various modeling techniques (including cluster analysis, CHAID, CART, discriminant analysis, latent class analysis, and Bayesian approaches) have been utilized. But the difference between segments is often captured by certain parameters in the model, which we believe is oversimplified. Their research scope is also limited by focusing heavily on brand choice and customer response to price, promotion and other marketing stimuli. The characteristics of customers used for segmentation are often represented with variables. We believe that customers' behavioral patterns are as important when analyzing the difference between customers.

In statistics and econometrics, some popular models split the input space (instead of objects to be clustered) according to the observed input variables, and a regression model is fit in each subspace. Examples of such models are the threshold autoregressive (TAR) model (Tong and Lim 1980), and Multivariate Adaptive Regression Splines (MARS) model (Friedman 1991). These models all decide where to split the input space according to the observed input variables. But often times, the driving force of splitting does not reside in the internal input variables. Gated Expert models (Weigend et al. 1995) introduce chosen external variables to detect the switching of regimes in time series data. It consists of a gating neural network and several competing neural networks. The gating network learns to predict the probability of the prediction of each expert. The input of the gating network includes chosen external variables which are picked manually. Under the situation that the driving force behind the splitting of the input space is unknown, it's not guaranteed that the hand-picked external variables will cover the hidden driving factors.

Another related stream is ensemble learning, where a set of classifiers is built, and then new data points are classified by taking a weighted vote of their predictions. Some well-known ensemble methods include bagging, cross-validation and boosting. Dietterich(2000) reviews these methods and explains why ensembles can often perform better than any single classifier. The fundamental difference between our approach and that of the ensemble methods is that, in their approach, no systematic difference among the data points are assumed and the sub-training sets are constructed more or less randomly.

In signature discovery and profiling community, studies have been mainly focusing on extracting features (variables) (Cortes et al. 2000) and generating rules (Adomavicius and Tuzhilin 2001) to represent signatures for an individual customer. Signatures are often used for personalization and fraud detection (Chan et al. 1999, Cortes et al. 2000). A key issue in developing personalization applications is constructing accurate and comprehensive customer signatures based on the collected data. However, as mentioned previously, there is no prior work which discovers multiple signatures for a customer through unsupervised learning and uses signatures to select local models as proposed in this dissertation.

One sub-problem within our research framework is clustering according to behavioral patterns. There has been vast amount of clustering research, but very few on clustering according to patterns which can be represented by items, itemsets, rules etc. In recent studies, several new clustering algorithms using items, itemsets and association rules (Agrawal et al. 1995) are proposed (Han et al. 1997, Wang et al. 1999).Compared to them, our method takes a new perspective of associating the representations (itemsets, rules etc.) with customers' behavioral patterns, and uses that concept to guide the clustering process. Wang et al. (2002) uses pattern similarity to cluster. They consider two objects similar if they exhibit a coherent pattern on a subset of dimensions. The definition of a pattern is defined as the correlation between attributes of objects to be clustered. This specific definition of pattern makes it more suitable for numerical data. But this type of pattern can certainly be adopted as another possible behavioral pattern representation for suitable applications.

Research Methodology

(a) *Hypotheses*

For datasets exhibiting heterogeneous patterns, we have the following hypotheses:

H1: Our approach outperforms the global model approach.
H2: Our approach outperforms the ensemble approach.
H3: Our approach outperforms other divide-conquer approaches.

(b) *Hypothesis testing methodology*

We propose to do two sets of experiments. The first set will evaluate if our grouping of transactions using patterns is "good". However, this is a hard problem since it is unsupervised learning. As a proxy, we propose to combine transactions from multiple users and test if our clustering method can separate transactions associated with each user. We have implemented the pattern-oriented clustering method using itemsets on user-centric web browsing data. User-centric web transaction data is data collected at the user level and thus captures entire history of web surfing behavior for each user. On one data set containing approximately 6000 transactions from two users, 95% of the clusters generated by our pattern-oriented clustering method are at least 95% pure (at least 95% of the transactions in a cluster belong to one user), while only 60% of the clusters generated by k-means on the original transactions are at least 95% pure. We plan to do more experiment on more data sets containing transactions from various number of users.

The second set of experiments will evaluate our entire divide-conquer approach. This is to test the above three hypotheses. As of now, we plan to test on three representative data sets. One is web purchasing data (predict purchasing), one is health care data (diagnosis of diseases), and one is financial data (discover technical trading rules). The predictive models we choose to use are regression models, decision trees, and neural networks. For hypothesis H2, we'll choose AdaBoost with trees (considered as the "best off-the-self classifier in the world" (Breiman 1998)). And for hypothesis H3, we'll choose one of the latest market segmentation models on one of the data set used for their model. Comparison will be conducted on both in-sample fit and out-sample prediction.

Expected Contributions

We believe that our approach will have the following contributions:

(a) *Deeper understanding of customer behavior*

As mentioned in the introduction, the existence of different categories of customer behavior is pervasive among business applications. By realizing the existence of the natural categories within customers' transactions and trying to discover those natural categories, we empower firms to understand their customers better and build more accurate customer models.

Theoretical approaches used in traditional market segmentation research give us much insight about the existence of the heterogeneity. The richer and more flexible representation, and powerful techniques used in data mining research make it possible to discover more complex and deeply embedded differences. By recognizing and discovering the differences in customers' behavior patterns, our approach can better separate the natural categories and further build more accurate models.

In traditional market segmentation research, data describing consumer preferences is typically obtained through survey or household purchase histories, which yield very limited individual-level information. The amount of data available for drawing inference about any specific consumer is very small, although there may exist many consumers in a particular study. This makes it very difficult to build individual level models. In today's world, this shortcoming becomes more striking, since personalization which is based on individual-level models is getting more and more popular. Our method can be used for situations where there is vast amount of transaction data. And in the situation where an single customer can generate large amount of transaction data (e.g. web browsing, credit card purchase, cell phone calls), our approach can be utilized to discover multiple signatures of a single customer, and build more sophisticated individual-level models.

(b) *A new way of discovering and utilizing signatures*

We point out that there exists an *unsupervised learning* problem, that has not been recognized in the literature, that is crucial in learning signatures. We propose a novel "divide and conquer" approach for building accurate consumer behavior models. Our approach uses signatures not only for data description but also for model selection. We also introduce the concept of distance between two signatures, and distance between a transaction and a signature. The former is used for clustering, and the latter is used for model selection.

(c) *Pattern-oriented clustering*

As a component of our model-building approach, our pattern-oriented clustering method itself has contributions to the clustering research. It incorporates customers' behavioral patterns into the clustering process. Our principle is to probe how transactions are generated in the first place so that we can discover more natural categories of transaction. This will help generating more meaningful and explainable clusters.

(d) *Wide applications*

Models in market segmentation are restricted to certain area such as band choice and pricing. Our approach can be used as a more general model-building technique, which gives it wide range of applications. It not only can be used for analyzing customer transactions, but also can be applied to other fields in which groups of objects differ in observable patterns (e.g. stock, credit card usage, gene expression etc).

References

Adomavicius, G. and Tuzhilin, A., "Using Data Mining Methods to Build Customer Profiles", IEEE Computer, vol.34, num.2 (2001) 74-82.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. I., "Fast Discovery of Association Rules", Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press, 1995

Allenby, G. M. and Rossi, P. E., "Marketing Models of Consumer Heterogeneity", Journal of Econometrics 89(1999) 57-78.

Breiman, L., "Arcing Classifiers (with discussion)", Annals of Statistics 26: 801-849.

Chan, P., Fan, W., Prodromidis, A., and Stolfo, S., "Distributed data mining in credit card fraud detection" , IEEE Intelligent Systems, 67-74, Nov/Dec 1999

Cortes, C., Fisher, K., Pregibon, D. and Rogers, A., "HANCOCK: A Language for Extracting Signatures from Data Streams", In Proceedings of the Sixth ACM International Conference on Knowledge Discovery and Data Mining, 9-17, 2000.

Dietterich, T.G., "Ensemble methods in machine learning", In J. Kittler and F. Roli (Ed.) First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science (pages 1-15). New York: Springer Verlag, 2000.

Friedman, J. H., "Multivariate Adaptive Regression Splines", Annals of Statistics, Vol.19, Issue 1, 1-67, March 1991

Han, E., Karypis, G., Kumar, V. and Mobasher, B., "Clustering based on association rule hypergraphs", in Proceedings of the SIGMOD'97 Workshop on Research Issues in Data Mining and Knowledge Discovery. 1997

Tong, H. , Lim, K.S. "Threshold Autoregression, Limit Cycles and Cyclical Data", Journal of the Royal Statistical Society, Series B(Methodological), Volume42, Issue3, 245-292, 1980

Wang, H., Yang, J., Wang, W., and Yu, P.S., "Clustering by Pattern Similarity in Large Data Sets", Proc. ACM SIGMOD Conference, Madison, WI, June 2002.

Wang, K., Xu, C. and Liu, B. "Clustering Transactions Using Large Items", Proc. 8th Int. Conf. on Information and Knowledge Management, Kansas City, November, 1999

Weigend, A. S., Mangeas, M. and Srivastava, A. N., "Nonlinear gated experts for time series: discovering regimes and avoiding overfitting", International Journal of Neural Systems 6, 373-399, 1995

Yang, Y. and Padmanabhan, B., "Pattern-Oriented Clustering of Web Transactions", Proc. 2003 Americas Conference on Information Systems (AMCIS 2003), Tampa, FL, August, 2003