iCARE: A framework for big data-based banking customer analytics

The amount of data stored by banks is rapidly increasing and provides the opportunity for banks to conduct predictive analytics and enhance its businesses. However, data scientists are facing large challenges, handling the massive amount of data efficiently and generating insights with real business value. In this paper, the Intelligent Customer Analytics for Recognition and Exploration (iCARE) framework is presented to analyze banking customer behaviors from banking big data, through analytical modeling methodologies and techniques designed for a key business scenario. Combining IBM software platforms and big data processing power with customized data analytical models, the iCARE solution provides deeper customer insights to satisfy a bank’s specific business need and data environment. The advantages of the iCARE framework have been confirmed in a real case study of a bank in southeast China. In this case, iCARE helps generate insights for active customers based on their transaction behavior, using close to 20 terabytes of data.

Introduction
The era of big data has arrived [1–3]. As discussed in [4], big data is being generated by a vast range of devices and processes. For example, numerous digital process and social media exchanges produce data trails. Systems, sensors and mobile devices transmit data. Big data is arriving from multiple sources with an alarming velocity, volume, and variety. Every day 2.5 quintillion bytes of data are created and 90% of the data in the world today was produced within the past two years [5]. In this big data era, the amount of data stored by any bank is fast expanding, and the nature of the data has become more complex. These trends provide a huge opportunity for a bank to enhance its businesses. Traditionally, banks have tried to extract information from a sample of its internal data and produced periodic reports to improve future decision making. Nowadays, with the availability of vast amounts of structured and unstructured data from both internal and external sources, there is increased pressure and focus on obtaining an enterprise view of the customer in a systematic way. This further enables a bank to conduct large-scale customer experience analytics and gain deeper insights for customers, channels, and the entire market. Integrating predictive analytics with automatic decision making, a bank can better understand the preference of its customers, identify customers with high spending potential, promote the right products to the right customers, improve customer experience, and drive revenue. However, data scientists are facing big challenges, including how to capture the massive amount of data in a cost-effective and efficient way, and how to sift through the big data to generate valuable business insights that translate into competitive advantages.

In this paper, the Intelligent Customer Analytics for Recognition and Exploration (iCARE) framework is presented as a method to efficiently analyze customer behavior using banking big data. Leveraging IBM products including IBM SPSS* Analytic Server and IBM InfoSphere BigInsights*, iCARE is able to process both unstructured and structured data and consolidate all the information to provide a unified customer view to yield new and deep insights into customer behavior. The iCARE analytical models are customized and validated on the processed data according to specified business scenarios, so that they can provide valuable business insights that could not be generated by traditional data mining models. The iCARE models are deployed in a parallel computing manner to achieve high performance and low response time. The solution can be personalized to satisfy a bank’s specific...
business need and data environment. The iCARE framework has been tested in a real case study involving a bank in southeast China. In this case, iCARE helped generate insights about the transaction behavior of active customers to develop data-driven marketing strategies.

The remainder of this paper is organized as follows. In the second section, some fundamental concepts are introduced, and then challenges in the big data era and the iCARE approach to solving the challenges are discussed. In the third section, the framework of iCARE is described in detail. Finally, the benefits of the iCARE framework are demonstrated from the results of a real case study in the fourth section. The conclusion is presented in the last section.

Background and motivation

Traditional banking customer behavior analytics

Today, with online banking and credit card and mobile payment systems, banks have access to a large amount of customer information. Adding to the complexity, the data is also increasingly available as “soft” information instead of “hard” information. For example, social media data can shed light on brand sentiment and brand loyalty. Hard data is usually recorded as numbers and easy to store and transmit in impersonal ways [6], whereas soft information is mostly communicated in text with the content dependent on the collection process. Thus, it becomes more difficult to process and analyze the soft information. However, it can help build a comprehensive understanding of the customer behaviors that can lead to new and important insights. For example, Agarwal et al. [7] noted the importance of soft information and suggested that it could be used in reducing overall portfolio credit losses. More recently, Lin et al. [8] investigated social networks and adverse selection in online peer-to-peer lending and found that the relational aspects of networks were consistently significant predictors of lending outcomes. Other authors [9, 10] empirically examined the effects of relationship banking in which soft information was fully utilized and proved the benefits in terms of customers’ default, attrition, and utilization behavior.

However, most traditional customer behavior analytic techniques are only focuses on hard information. As discussed elsewhere [11–14], traditional customer behavior analytics includes four dimensions: customer identification, customer attraction, customer retention, and customer development. Among them, customer identification is most fundamental and widely implemented in the banking industry. Customer identification includes customer segmentation and targeting. After customer identification, companies adopt appropriate marketing strategies to attract specific customers. An important element of the customer behavior analytics is improving customer retention and identifying the cause of customer attrition. The last dimension includes customer lifetime value analysis, up-selling/cross-selling, and affinity analysis for consistent expansion of transaction intensity, transaction value, and individual customer profitability.

Each component of customer behavior analytics mentioned above is linked to some standard data mining techniques. In customer identification, classification and clustering methods are usually used to target a specific customer group based on the business objective. In addition, regression techniques are applied to predict new potential customers. Almost all frequently used data mining techniques can be applied to better understand customer loyalty, including classification, clustering, sequence discovery, association, and regression. Association techniques are often used in customer development in affinity analysis to find the relationship between different products that are bought by a given customer over his or her lifetime.

Numerous solutions have been developed and studies done on traditional customer behavior analytics. For example, a framework for analyzing customer value and segmenting customers based on their value was proposed in [15]. Data on about 2,000 customers was used to train the segmentation model. Transaction history data of 150,000 credit card users were used in Hsieh and Chu’s paper [16], and the users were segmented into three different categories with different characteristics. Predicting customer turnover and high-risk customers in banking is presented in Prasad and Madhavi’s paper [17]. However, complexity of the data handled in most of the previous research is limited.

Challenges in the big data era

Compared with traditional customer behavior analytics methods most used in banking, there are two major challenges in the big data era. The first challenge involves how to handle the massive amount of complex data in a cost-effective and efficient way. The availability of data has grown in magnitude, speed, and dimensionality. As the number of channels that generate data has increased, so has the number of transactions and needed storage for related data. Moreover, valuable but unstructured log data is also collected from online banking system as a typical type of soft information. Traditionally, solutions to manage the large amount of data were unable to provide reasonable response times in handling expanding data volumes, leaving few options—either to run the analytics models on the large volume of data over days or perform piecemeal transactions for a more reasonable response time. Therefore, a bank needs to ensure the real-time or near real-time response for huge amount of data, which requires new expertise in the data management and the latest systems management methods. Additionally, new data analytical models are required to capture the value behind the increasing amount of unstructured, soft information.
The second challenge involves how to effectively generate business value from the analytics and obtain competitive advantages for banks. The most critical aspect of big data is its impact on how decisions are made. When data are scarce, expensive to obtain, or not available in digital format, it makes sense to let people make decisions directly based on experience or anecdotal evidence. In the big data era, a huge amount of information is created and transferred. Understanding of the problem should be combined with problem-solving techniques to improve decision making and bring real business value to a bank. According to McKinsey [18], trillions of dollars and Euros can be generated in value worldwide on the basis of big data. For example, $300 billion dollars in American healthcare, 250 billion Euros in European government, and more than 100 billion dollars in global personal location data services can be earned on an annual basis. Two charts in [18] are also used to compare the ease with which big data can be obtained and the value that can be expected from using big data in different sectors. From the charts, it is obvious that gaining data from the financial sector is relatively difficult due to the data volume and transaction intensity but has the potential to unlock new revenue streams.

The iCARE framework has been designed as a solution for these challenges. In the following sections, the iCARE framework is presented to analyze banking customer behaviors from banking big data through analytical models designed for key business scenarios. A real case of iCARE has been implemented that can generate insights for active customers on their transaction behaviors from dozens of terabytes of data.

The iCARE solution design

The architecture of the iCARE solution is demonstrated in Figure 1. There are four phases in the solution: data acquisition, data preparation, data modeling, and various business applications. Leveraging IBM software platform and data processing power, integrating data from multiple sources, and personalizing parallelized analytical models for different business applications, iCARE can provide deeper customer insights to benefit banks. Each phase is described in the following sections.

Data acquisition

With the development of new banking services, banks’ databases are evolving to adapt to business needs. As a result, these databases have become extremely complex. Since conventionally structured data is saved in tables, there is much opportunity for increased complexity; for example, a new table in a database is added for a new business or a new database replaces the old one for a business system upgrade. Aside from the internal data sources, there are structured data from external sources like economic, demographic and geographic data. To ensure the consistency and accuracy of the data, a standard input format is defined in iCARE for structured data. All the tables should be transformed to the format before input to the big data storage platform-the IBM InfoSphere BigInsights. This transformation procedure requires a thorough understanding of a bank’s business.

Furthermore, as mentioned, the growth of unstructured data creates even more complexity. While some unstructured data can originate from inside a bank, including web log files, call records, and video replays, more and more can be derived from external sources, such as social media data from Twitter**, Facebook**, and WeChat. The unstructured data is usually stored as files rather than database tables. Thousands of files with tens or hundreds of terabytes of information can be effectively managed on the BigInsights platform, which is an Apache Hadoop*-based, hardware-agnostic software platform that provides new ways of using diverse and large-scale data collections along with built-in analytical capabilities.

Data preparation

Because unstructured data is not organized in a well-defined manner, additional work must be done to transfer the data into a regularized or schematized form before modeling. The IBM SPSS Analytic Server (AS) provides big data analytics capabilities, including integrated support for unstructured predictive analytics from the Hadoop environment. It can be used to directly pull and query the data stored in BigInsights, eliminating the need to move data and enabling optimal performance on large amounts of data. Via tools provided by AS, methods for normalizing unstructured data can be designed and implemented on a regular schedule without writing complex code and scripts.
Even structured data needs additional data preparation to enhance the data quality on BigInsights with Big SQL (Structured Query Language), which is a tool provided by BigInsights as a combination of an SQL interface and parallel processing for handling big data. It can be used to efficiently handle the incomplete, incorrect, or irrelevant data. Furthermore, some statistical methods are implemented using Big SQL to reduce the impact of the noise in the data. For example, some data unreasonable values are detected and eliminated; some features are normalized or ranked. In this way, some highly suspected as outliers are controlled from hindering analysis. This step helps separate signal from the noise in big data analytics.

Once all the data has been prepared and cleansed, a data integration process is conducted on BigInsights. Data from multiple sources are merged, and the integrated data is stored in a data warehouse, in which the relationships between tables are well-defined, and data conflicts due to heterogeneous sources are resolved. Each full join between tables with millions of instances can be done on BigInsights in minutes, which usually takes hours without the parallel computing technique.

Based on the data warehouse, hundreds of attributes can be associated with each customer, and a consolidated enterprise customer view is generated. The iCARE analytical models will be built on the information.

**iCARE analytical models**

Based on the consolidated data, the iCARE analytical models can be built for different business scenarios based on business objectives. There are two advantages of the iCARE analytics models. First, all the statistic and machine learning methods in iCARE are customized to suit corresponding business scenarios. For example, when using the customer retention model, we developed a new, interactive decision tree, utilizing domain knowledge to optimize the model. Some possible business strategies will be generated automatically from the model to improve the data-driven decision-making process. The second advantage is that parallelized models are applied in iCARE to achieve better computing performance. For example, most data-mining algorithms need to scan the training data to determine model parameters. It requires intensive computing to access the data frequently, which is impossible to implement on a single-processor computer for large-scale data analytics. In order to improve the efficiency of algorithms, a parallel-programming method is needed [19]. For this purpose, all machine learning algorithms in iCARE are designed and developed to follow the MapReduce programming model. This parallel computing technique can save time and lower costs in model building and model evaluation. The next section gives an example of an analytical model built in iCARE to illustrate the advantages.

**An iCARE analytical model example: A customized and parallelized K-means clustering**

In this section, we will introduce a customized and parallelized K-means clustering algorithm as an example of the iCARE analytical models. The classic K-means clustering was proposed in MacQueen’s paper, which is an unsupervised machine learning technique used to divide \( n \) data points into \( K \) clusters (usually \( n \gg K \)) with similar points to minimize the total distance between the points to their cluster centers [20]. The algorithm first randomly selects \( K \) data points to be the cluster centers, then assigns each data point to the closest cluster, and updates the center of each cluster, usually using the average value of each dimension for all the data points in the cluster.

While simple, it is efficient and widely used in different applications to reduce the complexity and obtain initial insights on data. For example, it can be used to segment customers of a bank into several clusters based on their profile and transaction information, and then a bank can tailor its products, services, and marketing messages to each segment. However, there are several limitations when using it in large scale data analytics. First, usually the data quality is low. The clustering result is sensitive to errors and outliers, making it hard to interpret the business value of each cluster if the data is incomplete or inconsistent. Moreover, in customer data analytics, identifying a tight cluster with closely related data points is more valuable than dividing every data point into a cluster.

To improve the robustness of the K-means algorithm and generate segmentation result which is more appropriate for a bank, a customized algorithm has been developed as a part of the iCARE solution with the following steps:

1. Select \( K \) data points as the cluster centers.
   
   a) Use the first data point as the first cluster center.
   
   b) For each data point, compute the minimum distance between it and each defined cluster center. The Manhattan distance is used here. For two \( D \)-dimensional data points \( x \) and \( y \), the metric is defined as
   
   \[
   d(x, y) = \sum_{i=1}^{D} |x'_i - y'_i|, \tag{1}
   \]
   
   where \( x'_i \) is the coordinate of \( x \) in the \( i \)th dimension.
   
   c) Select the data point with the largest minimum distance from the defined cluster centers as the new cluster center.
   
   d) Repeat Steps b) and c) until \( K \) cluster centers have been added.

2. Assign each data point to the closest cluster using the standard K-means algorithm with distance metric shown in formula (1).
3. Update the cluster centers in the following way: suppose there are \( J_k \) data points \( x_{k,1}, \ldots, x_{k,J_k} \) in the \( k \)-th cluster where \( k = 1, \ldots, K \), and the current cluster center \( C_{k,\text{old}} \), then the updated cluster center is

\[
C_{k,\text{new}} = \sum_{j=1}^{J_k} w_{k,j} x_{k,j},
\]

where

\[
w_{k,j} = \frac{1}{d(x_{k,j}, C_{k,\text{old}})} \left/ \sum_{j=1}^{J_k} \frac{1}{d(x_{k,j}, C_{k,\text{old}})} \right.
\]

So the new cluster center is the weighted average of all the data points in a cluster, and the associated weight is inversely proportional to the distance between the data point and the old cluster center.

4. Redistribute the data points to their closest cluster and drop any data point \( x_j \), which is far away from any cluster center, that is

\[
\min_{1 \leq k \leq K} d(x_j, C_k) > \tau_1,
\]

where \( \tau_1 \) is a predefined distance threshold.

5. Apply (2) to update cluster centers using the remaining data points in each cluster.

6. Repeat Steps 4 and 5 until

\[
\max_{1 \leq k \leq K} d(C_{k,\text{old}}, C_{k,\text{new}}) < \tau_2,
\]

where \( \tau_2 \) is a predefined tolerance threshold.

This algorithm is further parallelized to follow a MapReduce model: the data points are split into different subgroups, and the Manhattan distances between data points and cluster centers are calculated in each subgroup (by the Mapper). To update the cluster center, partial weighted sum of data points in each subgroup can be computed in parallel, and then the Reducer will add up the partial sums to get the new cluster center. In each subgroup, the distances between data points and cluster centers are updated by the Mapper in parallel, and data points far away from any center are dropped. The next cycle will follow the same MapReduce model on the remaining data points. This procedure iterates until convergence.

There are notable advantages of applying this customized and parallelized \( K \)-means algorithm in banking data analytics. Using Manhattan distance instead of the more commonly used Euclidean distance makes the clustering algorithm more robust because the distances along each axis are not squared, making some error in one dimension less likely to skew the all the clusters. This is useful, especially for big data analytics where there may be a lot of noise in the data. Moreover, dropping data points that are not close to any cluster center reduces the noise of non-target customers for a product or service promotion, which is beneficial in terms of robustness and computing time. Applying (2) makes each cluster center better represent the key data points that are typical in a cluster and clearly capture the characteristics of its cluster. Finally, using the parallelized algorithm following the MapReduce model speeds up the implementation of the analytics model to process massive amounts of real banking data.

**Business applications**

The goal of customer analytics is to create a deeper understanding of customers and their behavior to maximize their lifetime value to the company. Customer analytics can be applied to many applications, like customer marketing, credit scoring and approval, profitable credit card customer identification, high-risk loan applicant identification, payment default prediction, fraud detection, money laundering detection, etc. The following are five examples that have been built in the current iCARE solution.

1. **Customer segmentation and preference analysis**: This module produces fine-grained customer segmentations in which customers share similar preference for different sub-branches or market regions. Based on these results, banks can get deeper insights in their customer characteristics and preferences, so as to improve customer satisfaction and achieve precision marketing by personalizing banking products and services, as well as marketing messages.

2. **Potential customer identification**: This module helps banks identify potential high-revenue or loyal customers who are likely to become profitable to the bank but are currently not customers. With this method, banks can get a more complete and accurate target customer list for high-value customers, which can improve marketing efficiency and bring huge profits to the banks.

3. **Customer network analysis**: By understanding customer and product affinity through analysis of social media networks, customer network analysis can improve customer retention, cross-sell, and up-sell.

4. **Market potential analysis**: Using economic, demographic and geographic data, this module generates the spatial distribution for both existing customers and potential customers. With the market potential distribution map, banks can have a clear overview of the target customers’ locations, and identify the customer concentrating/lacking areas for investing/divesting, which will support the banks’ customer marketing and exploration.

5. **Channel allocation and operation optimization**: Based on the banks’ strategy and spatial distribution of customer resource, this module optimizes the configuration (i.e., location, type) and operations of service channels (i.e., retail branch or automated teller machine). Maximizing revenue, customer satisfaction, and reach
against costs can improve customer retention and attract new customers.

Due to its ability to integrate multiple data sources quickly with parallel computing, iCARE provides a flexible framework that can be applied to many other applications. In the next section, a specific case study and its results are examined.

A case study

To verify the iCARE framework, we have conducted a real case study with a commercial bank in southeast China. The bank wanted to transform customers from a traditional service retail channel to online retail in order to reduce operational costs. The higher the online banking customer active index (an index defined by the bank to evaluate how actively a customer uses the online channel for transactions), the lower is the pressure on conventional channel services. In this case, close to 20 terabytes of data was analyzed in iCARE to help generate insights for retaining active online banking customers, and identify the customers who were more likely to drop off based on transactional behavior. Based on information gleaned about these customers, personalized retention strategies would then be developed to maintain the customer active index. The framework can be found in Figure 2.

This bank had large amount of structured data with ambiguous definitions in their various systems, including the online banking system, E-payment platform, Enterprise Customer Information Facility (ECIF) system, and core banking system. Additionally, the data had multiple formats, corrections, and incompatible code types. Explanations of various code values were unclear, outdated, or difficult to extract. Understanding and transforming the data to the standard format required by iCARE required more than a month. After that, the data was loaded into BigInsights. The tool Big SQL was utilized to clean and prepare the data, including filling in the blank value with the most common value and detecting outliers with statistical methods.

In addition, the online/mobile banking log files were also provided for analysis. These log files were unstructured and had never been analyzed before. The log schema was carefully studied before the contents could be translated into structured data. After understanding the format and schema in which the log files were recorded, the team extracted the structured information using SPSS AS.

More than two hundred of attributes were generated from different sources, including personal information, account information, and transactions. The personal information included age, gender, job type and other demographic information. The account information includes

Figure 2

The framework of the iCARE solution in a real case.
the application date of the account, what type of business has been opened and the opening date, and other required application information. Recency and frequency of the transactions via each channel and online banking website are major components of the transaction data. It is worth mentioning that the online banking behavior attributes were obtained from the online banking log files by a customized information extraction algorithm executed on SPSS AS in a parallel manner. Then, the data from multiple database sources were integrated to form a unified view of the customer. At this point, issues with data inconsistency due to the heterogeneous sources were checked and resolved. For example, we wanted to merge the data from the online banking system and core banking system. However, the key variables in the two systems were different. Thus, we used customer ID to build a mapping between the online banking system and the ECIF system, and then used the card number to build another mapping between the core banking system and the ECIF system. Then, the mapping between online banking system and core banking system was established after analyzing the complex structure of ECIF system and eliminating some conflicting records.

After the data preparation, the models were built to identify active customers who had a high possibility of becoming inactive in the future. The performance of our model is shown in Figure 3. The precision is measured by the percentage of correctly classified customers. Compared with the baseline in which customers likely to become inactive are randomly chosen in the dataset, the customized decision tree significantly improves the percentage of correct identification. As shown in Figure 3, the precision of our model is approximately 1.59 times higher than the baseline result from random selection when a list of 30,000 customers is reported. The customized decision tree is comparable to other benchmark models in terms of performance and computing time, but can generate rules that are easier to understand and interpret. The output was a list of customers having a high possibility of becoming inactive, and the associated rules covering different perspectives of customer behavior. For example, a customer who was married, had an online banking transaction amount of less than 453 RMB (Renminbi, the official currency in China), logged in the online banking system for less than five times, and had an average login time of less than 481 seconds over the past two months would probably become inactive in the next month. Here, the average login time appeared in the rule since the online banking log had been analyzed. Furthermore, some potential retention strategies were generated from the model for each identified customer. The results helped the bank’s marketing team promote specially designed products or services to the target customers.

Our model ran parallel on SPSS AS, which was significantly faster than unit operation. Figure 4 is a test for comparison of computing time between the big data platform and a single host. The model ran 12 times faster as a single host for the 4 GB test data sample with 1,600 instances.

Conclusion
In this paper, the iCARE framework was introduced as a method to handle massive amount of data efficiently and
analyze customer behavior for retail banks. It goes beyond the limitation of traditional customer analytics that has been done in the banking industry, using unstructured data that has not been used before. Furthermore, the results can be interpreted as business rules that help with decision-making that can generate real business value for a bank. A real application of the iCARE framework implemented for a bank in China seeking to increase customer retention in their online banking channel was described in detail based on IBM products including SPSS AS, SPSS Modeler and BigInsights. The results of the case study demonstrate that the iCARE framework provides a useful way to handle complex banking data analytics.

However, this paper just outlines preliminary work on the iCARE framework; there are many possible extensions of the method. It can be extended for other data analytics applications, not limited to customer relationship management or the banking industry. Finally, since the iCARE framework is scalable by adding more parallel analytical models, this lays the ground work for even bigger projects using more data.

*Trademark, service mark, or registered trademark of International Business Machines Corporation in the United States, other countries, or both.

**Trademark, service mark, or registered trademark of Twitter, Inc., Facebook, Inc., or Apache Software Foundation in the United States, other countries, or both.

References


Received January 16, 2014; accepted for publication February 11, 2014

Ning Sun IBM Research - China, Beijing 100193, China (sunbh@cn.ibm.com). Dr. Sun is a Staff Researcher in the Department of Business Analytics and Optimization at IBM Research - China. She received a B.S. degree in information management and an M.A. degree in finance from Peking University. She received an M.A. degree in statistics from Columbia University and a Ph.D. degree from Stony Brook University. She subsequently joined IBM Research - China. Her research interests include statistics and machine learning.

Jacqueline G. Morris IBM Global Business Services, New York, NY 10016 USA (jgmorris@as.ibm.com). Ms. Morris is a senior consultant in the advanced analytics practice area of Business Analytics and Optimization at IBM Global Business Services. She received her B.A. degree in mathematics and economics from Columbia University. Her interests include customer and patient outcome analytics.

Jian Xu IBM Research - China, Beijing 100193 China (xujiao@cn.ibm.com). Mr. Xu is a Researcher in the Department of Business Analytics and Optimization at IBM Research - China. He received a B.S. degree in mathematics from Soochow University in 2010, and an M.S. degree in computational mathematics from Peking University in 2013. He subsequently joined IBM Research - China. He is author or coauthor of three technical papers. Currently, he is working on banking analytics.
Ms. Zhu is a Researcher in the Department of Business Analytics and Optimization at IBM Research - China. She received a B.S. degree in statistics from Hua Zhong University of Science and Technology in 2010 and an M.S. degree in statistics from Peking University in 2013. She subsequently joined IBM Research - China, where she has worked on the research of big time series data analytics. Her research interests include statistics and machine learning.

Dr. Xie is a Research Staff Member and manager of the Department of Business Analytics and Optimization at IBM Research - China. He received his B.S., M.S., and Ph.D. degrees in management science and engineering from Tsinghua University, China. He subsequently joined IBM Research - China, where he works on the research of advanced optimization and analytics and their applications in cross-industries. He has multiple patents and publications in refereed conferences and journals. His awards include the INFORMS (Institute for Operations Research and the Management) Franz Edelman Finalist Award, the IBM Research Accomplishment Award, IBM Outstanding Technical Achievement Award, and Beijing Municipal Science and Technology Award. His research interests include statistics, data mining, optimization, and simulation.