Building a Credit Rating System for Credit Card Issued Banks Using Data Mining Techniques and Cloud Computing

Chung C. Chang, Professor, Department of Information Management, Chinese Culture University, Taiwan
Vo Hoang Nha, M.S., Department of Information Management, Chinese Culture University, Taiwan

ABSTRACT

Data mining and cloud computing are two major trends in the recent information technology development. Data mining is a process to get actionable information and cloud computing provides software services via the Internet. One of the techniques which has been widely used in data mining for classification tasks is the nearest neighbor algorithm or NN algorithm. The NN algorithm is a method to classify objects based on the shortest distance among objects in the data base. In this paper, we use the NN algorithm to rank customers’ credit for credit card issued banks. Besides that, we use cloud computing as a platform to provide a simple, convenient and low-cost way for bank users in credit rating.

Keywords: credit rating system, data mining, nearest neighbor algorithm, cloud computing

INTRODUCTION

Credit rating is widely used in business. For example, when a company receives an order from a customer, it must first validate the customer’s credit status before shipping the goods. For a credit card issued bank, the credit rating operation is even more important since the credit card business itself involves high potential risk. Traditionally, banks rely on some heuristic rules based on their own expertise to evaluate each credit card applicant in order to raise the profit as well as to reduce the risk. However, due to each credit card applicant is different in age, gender, occupation, income, and so forth, a one size fit all rule is difficult to find to effectively deal with this situation.

Instead of using rules or rule-based systems, this paper uses a data mining technique called the nearest neighbor algorithm to propose a better and more practical way to solve the problem. By using the NN algorithm, the system can learn from the past cases in which the results are known and then apply them to the current problem at hand. In other words, we may find out the most similar cases or the nearest neighbors in the data base in order to find the answer. From the information technology point of view, the similarity is measured based on the distance between any case in the data base and the current case.

Cloud computing is one of the major trends in information technology development today. It is the concept of using the Internet as a medium to transfer all data stored on a computer to a remote server for storage or processing. Cloud refers to the data center consisting of numerous servers. When you connect to these servers on the Internet, the computer will process data for you (Chang and Liu, 2010; Guo, Chen, Chen and Tang, 2010). This study combines data mining techniques with cloud computing, building the credit rating system on a cloud computing platform to provide a simple, convenient and low-cost way for bank users.
LITERATURE REVIEW

Data Mining and the Nearest Neighbor Algorithm

Data mining was defined by Kamala (2013) as the process that attempts to discover patterns in large data sets. The overall goal of the data mining process is to extract information from a large data set and convert it into an understandable structure for future use. Data mining is the process of discovering or finding new, valid, understandable and potentially useful forms of data. Nikam and Patil (2013) also indicated that data mining is the process of finding correlations or patterns among fields in large data sets and building up the knowledge-base based on the given constraints. This process is often referred to as knowledge discovery in data base (KDD). It encompasses data storage and access, scaling algorithms to very large data sets and interpreting results. The data cleaning and integration process included in data warehousing facilitate the KDD process.

Han and Kamber (2001) explained that KDD process consists of seven major steps as follows:
1. Data cleaning: To remove noise and inconsistent data.
2. Data integration: Where multiple data sources may be combined.
3. Data selection: Where data relevant to the analysis task are retrieved from the data base.
4. Data transformation: Where data are transformed or consolidated into forms appropriate for mining.
5. Data mining: An essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation: To identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation: Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

The nearest neighbor algorithm, or NN algorithm, is one of the commonly used data mining techniques. It is often used with the concept of memory-based reasoning in data mining for finding the most similar data or cases in a data set. The concept of memory-based reasoning attempts to find out the answer for the current case from the similar cases in the past which answers are known. Peterson (2009) defined the nearest neighbor algorithm as a classification method basic and simple and should be one of the first choices for a classification task when there is little or no knowledge of the distribution of the data. Classification of the nearest-neighbor was developed from the need to perform separate analyzes when reliable estimates of the probability density parameter is unknown or difficult to determine.

Imandoust and Bolandraftar (2013) indicated that the nearest neighbor algorithm has several main advantages: simplicity, efficiency, intuitiveness and competitive classification performance in many areas. It is robust to noisy training data and effectively if the training data is large. The NN algorithm also has several limitations. It may have poor run-time performance when the training data set is large. It is very sensitive to the features irrelevant or redundant because all these features contribute to the same and thus to classify. By carefully selected features or essential features, this can be avoided.

Cloud Computing

The cloud is an environment where users only need an application or browser to access services through the Internet. Users are neither required to understand the arrangement of the cloud, nor download or update the system, making services easily accessible and scalable. Mell and Grance (2011) pointed out that cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.
The features of cloud computing described by Stanoevska-Slabeva and Wozniak (2009) are:

• Cloud computing is seen as a new computing paradigm.
• The infrastructure and resources (hardware, storage, system software and applications) are provided on the X-as-a-Service basis. When these services are provided by the vendor of cloud computing, customers want to use them must pay fees.
• All the features of cloud computing is virtualization; it is easy to expand as needed.
• Computer gadgets and SaaS are provided in an integrated way, although utility computing can be consumed separately.
• Cloud services are provided through a web browser or through a defined API.

Kuyoro, Ibikunle and Awodele (2011) described that in the model of cloud computing deployment, networking, storage, and software infrastructure are provided as services that scale of the service up or down depending on the demand. The cloud computing model has three main deployment models: the private, public and hybrid cloud. In the private cloud model, infrastructure and services only built to serve an organization (enterprise). This enables enterprises can control a maximum of data, securely and effectively. Private cloud can be built and managed by the IT team of business or can hire a service provider.

Public cloud provides services for people widely used. These services are provided and managed by suppliers and applications are on cloud system. Service users will benefit from the lowest investment costs and reduced risk. The suppliers must be responsible on managing the system, infrastructure, security, and so forth. Hybrid cloud is a private cloud linked to one or more external cloud services. Public cloud is easy to apply, low cost but not safe. In contrast, private cloud safer but costly and difficult to apply. Thus, if combined these two models together, it will exploit the advantages of each model. That is the idea of forming hybrid cloud model (Kuyoro, Ibikunle and Awodele, 2011; Stanoevska-Slabeva and Wozniak, 2009; Mell and Grance, 2011; Liao and Chang, 2016).

Whether to choose the private, public or hybrid cloud basically depends on the enterprise specific needs. Since different cloud computing model has its advantages and disadvantage, enterprises should consider for the cloud computing model that they choose to deal with different applications. Generally speaking, private or hybrid cloud may fit business permanent needs; on the other hand, public cloud may be more suitable for business temporary needs.

THE CREDIT RATING SYSTEM

Algorithm Used in the System

The algorithm used in the system to find out the K nearest neighbors is as follows:
1. Determine the features of each case or applicant such as age, gender, income, etc.
2. Use historical cases to set up a training data set for the credit rating system.
3. Determine the number of categories. In the credit rating system, there are three categories of credit rating: A represents excellent, B stands for fair and C means poor.
4. Determine the value of K.
5. Calculate the distance between the case in hand and each case in the training data set. The Euclidean distance function is used for calculating the distance between case A and B is as follows and where D stands for the distance function and the number 1 to n represents the different features of case A and B.
\[ D_{euclid}(A,B) = \sqrt{D_1(A,B)^2 + D_2(A,B)^2 + ... + D_n(A,B)^2} \] (1)
6. Sort the distances from minimum to maximum and select the K nearest neighbors with minimum distances.
7. Since the results in the K nearest neighbors have already known, we can easily find out the result and the confidence for the current case by using a majority voting.

**The System Prototype**

A system prototype of the credit rating system has been built in order to examine and verify the research ideas. This system prototype consists of three main steps or functions: user input customer features, process data (use NN algorithm) and output results as shown in figure 1. In the first step, the user input the features of the current customer such as gender, age, income, number of dependents, with or without assets, etc. in order to look for the most similar cases or neighbors. Secondly, the system will use the NN algorithm to find out the K nearest neighbors based on the similarity between the current case and the cases stored in the data base. Finally, the system will generate the credit rating results for the credit card issued bank.

![Figure 1: The three major steps or functions in the credit rating system](image)

Based on the input requirements of credit rating for individual customer, we build a table of customer including data fields and a partial training data set as shown in figure 2 and figure 3. A sample result generated by the credit rating system is shown in table 1. Figure 4 shows the credit rating system works in a cloud computing environment.

![Figure 2: Data fields in the table of customer](image)
Figure 3: Partial training data in the credit rating system

Table 1: A sample result generated by the credit rating system

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Vo Minh Phu</td>
</tr>
<tr>
<td>Dependent</td>
<td>2</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
</tr>
<tr>
<td>Age</td>
<td>32</td>
</tr>
<tr>
<td>Asset</td>
<td>Yes</td>
</tr>
<tr>
<td>Value of K</td>
<td>4</td>
</tr>
<tr>
<td>Income</td>
<td>78000 USD/Year</td>
</tr>
<tr>
<td>Credit rating</td>
<td>Excellent - Confidence: 100% (AAAA)</td>
</tr>
<tr>
<td>Advice</td>
<td>Low risk, should give highest credit limit</td>
</tr>
</tbody>
</table>
CONCLUSION

Data mining and cloud computing are two major trends in the recent information technology development. Data mining is a process to get actionable information and cloud computing provides software services via the Internet. One of the techniques which has been widely used in data mining for classification tasks is the nearest neighbor algorithm or NN algorithm. The NN algorithm is a method to classify objects based on the shortest distance among objects in the data base. In this paper, we use the NN algorithm to rank customers’ credit for credit card issued banks. Credit rating system is an important tool in risk management for banks’ credit rating activities. The results of the credit rating system are used for credit card applicants in deciding their credit rating. The training data set used for building the system is from a bank in Vietnam. Although the system still has some limitations such as slow processing speed, they may be solved in the future research works. Besides that, integrating cloud computing with the credit rating system helps reduce operating costs and provide a simple and convenient way for the bank users in customer credit rating.

REFERENCES