Context-based Knowledge Discovery and Its Application

Lingling Zhang1,2*, Xiao Wang1, Liang Zhang1, Yibing Chen1,2 and Yong Shi1,2,3

1School of Management, Graduate University of Chinese Academy of Sciences, Beijing 100190, China
2Research Centre on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China
3College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182, USA

ABSTRACT
Mankind is inundated by information, but thirst for knowledge. The use of knowledge discovery to identify potentially useful knowledge from massive data has become an important method, which increasingly attracts much attention. In order to solve the problem of too much emphasis on the accuracy of the algorithm while ignoring the context of the application of knowledge existing in traditional knowledge discovery, we proposed theoretical framework of context-based knowledge discovery. Through the study of context representation based on probability distribution and calculation of context variance and distance etc, data selection based on similarity assessment of context is achieved. Further a context-based KNN classification algorithm is designed. Finally the validity of context-based knowledge discovery is verified.

Key words
context; knowledge discovery; data selection; KNN; knowledge reuse

1. INTRODUCTION
Mankind is inundated by information, but thirst for knowledge (Naisbitt, 1982). Advances in computer hardware provide large-capacity storage devices. Development of database provides the possibility of the storage of exponentially growing amount of data. Extensive use of IT technology makes sure policy-makers can get vast amounts of data, which has exceeded the limits of human processing capacity (Jiawei Han, Micheline Kamber, 2007).

As an important part of today's research in the field of intelligent systems, knowledge discovery can obtain innovative, interesting, potentially useful and ultimately understandable knowledge depending on powerful computing capability (Fayyed et al. Extracted from the vast amounts of data, 1996). However, significant deficiencies of traditional knowledge discovery leads to deviation between knowledge extracted and practical needs, which is difficult for decision making and popularization. Detaching from context is one of the causes.

This is because in practice, the vast majority of knowledge discovery methods is driven by data and technology, and ignores context factors which are closely related to the knowledge discovery process.

First, knowledge comes from variety of context and it cannot be accurately understood without context (Brezillion, 1999). If knowledge separates from other relevant knowledge, limitations and distortions of understanding will appear (Brezillion, 1999)(Goldkuhl, 2001)(Klemke, 2000). In some contexts, knowledge is correct but may be wrong in others, such as "winter is peak season of down jacket; sales will increase" is suitable in northern enterprises, but not adaptable in southern ones where there exist no temperature differences. If we use the knowledge regardless of context, a substantial increase in the down jacket supply will only lead to the hoarding of goods and huge loss.

Second, context is the perception of things around (Despres, 2000). Context and personality are very important in knowledge management (Dierig, 1999). Knowledge reflects a particular stance, perspective, or intention in accordance with the characteristics of specific context, which is different from information (Nonaka, 1998). The value of knowledge depends on its context (Cohen, 1998) and plays a role among the specific context (Thompson, 2004). All knowledge would be meaningless without context. Because corporate culture, organizational structure and business processes are different, each corporate is different and it is difficult to copy the experience of knowledge management in other enterprises. Amending knowledge in accordance with the specific context of the enterprise and managing knowledge in a specific context ensures its effectiveness in the new enterprises.

2. LITERATURE REVIEW
In recent years, domain-based knowledge discovery has gradually been concerned about (Domingos, 2007) (Graco, 2007). How to add domain knowledge into the knowledge discovery process have increasingly become the hotspots of the research at home and abroad. From the perspective of knowledge discovery, domain knowledge refers to the guidance and constraints joined into the knowledge discovery process to dig out the collection of constraints of the knowledge interesting to users (Maja and Theo, 1999). However, the definition of domain knowledge and context is not clear. So we often refer domain knowledge to expertise, context and other factors.

The similarity of context and domain knowledge is used to constrain the knowledge discovery process, so as to get knowledge interesting to users. However, the scope of domain knowledge is more extensive, including expertise, user preferences and other human factors. Context refers to objective influence factors and it can be regarded as a specific case with its unique applications. They cannot be generalized. For example, Sinha AP and Zhao H (2008) captured the knowledge of borrowers’ credit level determined by loan experts and added it into data mining algorithm. Guo Jiansheng (2001) conceptualized the value of continuous attributes and automatically partitioned clustering data. Domain knowledge here especially refers to the experience of experts in the field of domain knowledge.

In this sense, context is also a kind of domain knowledge. Specifically, context can be regarded as a subset of the domain...
knowledge. Part of the study of domain knowledge refers to expertise. Shu Fengdi (2007) provides personalized domain knowledge for the demand of application software based on user characteristics and context. Warwick Graco (2007) proposed a knowledge-oriented data mining which should consider expertise, smart data and business knowledge in addition to the algorithm. Data mining should become an intelligent process. Combining with technical knowledge, business knowledge here refers to the context of this study. Based on different functions, the accuracy rate is higher than traditional data mining (Midelfart, 2001). Yang Li proposed an in-depth study should be conducted, which is about how to combine with the knowledge discovery algorithms, and how to adapt to the distributed, collaborative environment (Yang Li, 2005). In the process of data preprocessing, Kuo et al (2007) applied medical ontology to classification attributes of association rule mining and the expression of mined knowledge has been improved. Zhu Jingbo, Chen Wenliang (2005) used associated words to express text characteristics and proposed a new clustering formula, which effectively improved the performance of text classification. Zhu Zhengxiang and Gu Jifa (2009) believed that traditional data mining algorithms are inadequate in practice, and domain knowledge should be added into data mining process. Chen Shengqing (2009) proposed knowledge discovery process model combined with domain knowledge discovery, providing reference for the design of knowledge system model with no explicit concept of knowledge context. Combining with technical knowledge, business knowledge here be conducted, which is about how to combine with the context concept and principles. Separating context from knowledge discovery is not much, and there are following defects:

(1) There are discussions about the integration of context and knowledge discovery in theory and practice. But research about the combination of context and knowledge discovery is not much, and there are following defects:

(2) Currently most research about context and knowledge discovery is to solve a specific problem, or to combine a certain stage of knowledge discovery and certain context attributes. There is lack of systemic context concept and principles. Separating context from domain knowledge and studying its guiding role to knowledge discovery are necessary.

(3) The in-depth study about how to add context to knowledge discovery is still the focus of further research.

3. CONTEXT-BASED SIMILARITY ASSESSMENT METHOD

In order to combine context and knowledge discovery, the critical research is to find method of representing context. Depending on the characteristics of the dataset, we proposed context representation based on probability distribution. This method considers the information of statistical category attribution and context, which transforms context factors into probability distribution. In this way, context can be well expressed as numerical values. On the one hand, context information is kept and added into the result of knowledge discovery; on the other hand, the concept is expressed as the numerical description, which will be conducive to the combination of knowledge discovery.

3.1 Context Representation Based on Probability Distribution

The function of knowledge discovery is to find new rules from large amounts of data. The basis of model construction and decision making is data. Assuming D is training set from the dataset A. Supposing that the class label has m different values and there are m different classes L1, L2, ..., Lm. Li is a collection of class L1 in D, |D| and |Li| are the number of tuples of D and Li.

Tuples \( D_1^k \) in D satisfy \( C^j = C^k_j \) \( (k = 1, 2, \ldots, s) \), that is, \( D^k_j = D_1^k \). Depending on different class labels, it can be divided into m subsets \( \{L_{k,1}, L_{k,2}, \ldots, L_{k,m}\} \):

Let \( L_{k,i} \) is tuple set of class \( L_i \) in \( D^k \), that is, tuples satisfy \( C^i = C^k_i \) and their class label is \( L_i \).

\[ |D^k_i| \text{ and } |L_{k,i}| \text{ are on behalf of the number of tuples } D^k_i \text{ and } L_{k,i}. \]

For any \( C^j (k = 1, 2, \ldots, s) \), existing \( (p_{L_{j,k}}, p_{L_{j,k}}, \ldots, p_{L_{j,k}}, \ldots, p_{L_{j,k}}) \) on behalf of the corresponding probability distribution of a particular context attribute.

\( p_{L_{j,k}} \) is the probability of tuples belonging to the class \( L_j \) in \( D^k \), that is the proportion of tuples satisfying \( C^j = C^k_j \) and category \( L_i \) in D and \( D^k_j \), which can be calculated by

\[ \frac{|L_{k,i}|}{|D^k_i|} \]

\[ p_{L_{j,k}} = \frac{|L_{k,i}|}{|D^k_i|} = \frac{|D(C^j = C^k_j \cap L_i)|}{|D(C^j = C^k_j)|} \]

For example, \( |D| = 1000 \), they can be divided into \( L_1 \), \( L_2 \), \( L_3 (m = 3) \) and the context attribute \( C^j \) has two values \( C^j_1 \), \( C^j_2 (v = 2) \). Supposing the dataset satisfying \( C^j = C^j_1 \), \( |D^j_1| = 300 \), of which 100 data belongs to \( L_1 \), 50 data belongs to \( L_2 \) and 150 data belongs to \( L_3 \), the probability distribution of context attribute \( C^j_1 \) is \( (1/3, 1/6, 1/2) \). The first vector represents the probability belonging to class \( L_1 \) divided.
by the total number). Similarly assuming the probability distribution of context attribute $C_i^j$ is $(2/7, 1/7, 4/7)$.

### 3.2 Data Selection and Weight Determination Based on Context Variance

Through the context representation based on probability distribution, we get different probability distributions of values of various context attributes. Supposing there are $n$ context attributes, any value of the context attribute can be expressed as $1 \times m$ vector, where $m$ is the number of classes. For datasets containing $q$ data, each context attribute corresponds to the $q \times m$ vector. These $q$ data corresponds to $s$ different values of context attributes.

Every context attribute has several values and each value can be represented in the form of probability distributions. In order to reflect the importance of different context attributes, it is necessary to apply the degree of dispersion fluctuations to reflect the importance indirectly. The basic idea is to calculate the variance of context attributes.

Context variance of different context attributes reflects the distinction of knowledge discovery results. The variance can be used for the selection of context attributes and weight determination. On the basis of context representation based on probability distribution, we can select context attributes and determine weights through the calculation of variance based on the probability distribution.

#### 3.2.1 Calculation of Context Variance

The mean and variance of context attributes can be calculated through the context representation based on probability distribution, which can be calculated as follows:

For any data $t (t = 1, 2, \ldots, q)$, the context attribute $C_i^j$ has the value

$$ P_t^j = \left( p_{i_1,k,t}^j, p_{i_2,k,t}^j, \ldots, p_{i_m,k,t}^j \right) \quad (i = 1, \ldots, m; \ j = 1, 2, \ldots, n)$$

Then

$$ P = \frac{1}{q} \left( \sum_{t=1}^{q} P_{i_1,k,t}^j, \sum_{t=1}^{q} P_{i_2,k,t}^j, \ldots, \sum_{t=1}^{q} P_{i_m,k,t}^j \right) $$

Further, the variance of context attribute $C_i^j$

$$ \text{Var}^j = \frac{1}{q-1} \sum_{t=1}^{q} (P_t^j - \bar{P})^T (P_t^j - \bar{P}) $$

$\text{Var}^j$ denotes the dispersion level of values of context attribute $C_i^j$. The greater context variance is, the distinguishing ability of context attribute is bigger.

Dataset $|D_1^j| = 300$ satisfying $C_i^j = C_i^j$ , $|D_2^j| = 700$ satisfying $C_i^j = C_i^j$, the probability distribution of context attribute $C_i^j$ is $(1/3, 1/6, 1/2)$ and the probability distribution of context attribute $C_i^j$ is $(2/7, 1/7, 4/7)$ then

$$ \bar{P} = \frac{1}{1000} \left( \frac{1}{3} \times 300 + \frac{2}{7} \times 700, \frac{1}{6} \times 300 + \frac{1}{7} \times 700, \frac{1}{2} \times 300 + \frac{3}{7} \times 700 \right) $$

$$ \text{Var}^j = \frac{1}{1000 - 1} \times \left[ 300 \times \left( \frac{1}{3} - \frac{3}{10} \right)^2 + 700 \times \left( \frac{2}{7} - \frac{20}{3} \right)^2 \right] $$

$$ \text{Var}^j = 0.001 $$

However, not all context attributes are useful to knowledge discovery. Redundant context attributes may interfere with mining results and reduce valuable context effects. In order to eliminate the effects, it is necessary to take a reasonable approach to select practical and useful context attributes.

#### 3.2.2 Context Selection Based on Similarity

Context is involved with environment, background, goal, emotion and other factors, which contain complicated context elements. If all context factors are joined into knowledge discovery, the effect of the model cannot only be improved but also be negative. Thus, we should select useful context factors collected to guide the whole process of knowledge discovery.

Context variance can be used to indicate the discrete situation of the data belonging to different classes under the role of context factors. If context variance is big, the data tend to be distributed to different classes which means context information gain is big and the distinction between different dataset. Conversely, if the calculated variance is small, the distinction role is not significant.

Context variance reflects the contribution to the classification results. We tend to choose attributes with larger variance as influencing factors to be considered. Through the calculation of corresponding variance of the context, we sort reversely by the variance and choose the first $N$ context attributes so as to compare context effect of each attribute.
3.2.3 Determination of Context Weight

Through the selection of context attributes by removing useless redundant attributes, the N attributes have guiding roles to mining results. But different context attributes have different influencing effects to knowledge discovery results. In order to reflect the importance of differences and improve the efficiency of knowledge discovery process combined context, we use weights while applying context factors in the knowledge discovery process.

In order to ensure the sum of weights is equal to 1, normalized calculation of context variance is as follows,

$$w_j = \frac{\text{Var}_j}{\sum_{j=1}^{N} \text{Var}_j}$$

It is necessary to give a certain weight to some context factor to express different distinguishing degrees of different context. We can refer context variance above \(\text{Var}_j\) (\(i = 1, \ldots, N\)), that is \(w_j = \text{Var}_j\).

\(w_j\) is used to represent the weight of each context attribute, which is a reflection of importance of different context. Here, \(\sum_{j=1}^{N} w_j = 1\). \(\text{Var}_j\) stands for the variance of context attribute \(j\).

Further we can calculate context similarity through context attribute and weight. Data selection and knowledge reuse can be improved through identification and evaluation of similar context.

3.3 Context Similarity Assessment Methods

An important technology of combining context and knowledge discovery is examining context similarity of different data. Next the method of computing context similarity provides a new way of thinking to data selection and knowledge reuse.

Any data has not only natural property for traditional knowledge discovery, but also different context attributes separately. In practice, due to the difference of context attributes, rules obtained cannot be applied to decision-making. To solve this problem, the study of context similarity is the key point.

3.3.1 Calculation of Single Context Similarity

Context factors of one object are combination of a series of context attributes. We first calculate single context distance, which is context difference. Through context representation based on the probability distribution, the value of context attributes can be expressed as \(p_{1j,k}, p_{2j,k}, \ldots, p_{nj,k}\). The “distance” of context attributes of every two data can be calculated as follows: Europe Euclidean distance of each context value based on the probability. The distance shows proximity between the two data points, or the dissimilarity. The distance of certain context attributes belonging to different data points can be compared. Specific calculation is as follows:

For any data \(i(t = 1, 2, \ldots, q)\), the value of context attribute \(C_i\) is

$$p_{ij} = (\frac{p_{i1,k}, p_{i2,k}, \ldots, p_{ij,k}}{\sum_{j=1}^{N}}) (j = 1, 2, \ldots, N; k = 1, 2, \ldots, s)$$

The difference between \(x\) and \(y\) is

$$d_{x,y} = \sqrt{\sum_{j=1}^{N} (p_{ix,k} - p_{iy,k})^2}$$

\(d_{x,y}\) represents the dissimilarity in a single context attribute \(C_i\) between object x and y. The more similar two objects are, the closer to 0 the value is. The more different two objects are, the bigger the value is.

Continuation of the example 4.1, assuming that the value of context attribute \(C_i\) of the first data is \(C_i\) and the value of context attribute \(C_i\) of the second data is \(C_i\). Then, the context distance \(d_{i,j}\) of the two data is as follows:

$$d_{i, j} = \sqrt{(\frac{p_{i1,k}, p_{i2,k} - p_{i1,k}, p_{i2,k}}{\sum_{j=1}^{N}})^2 + (\frac{p_{i2,k}, p_{i2,k} - p_{i2,k}, p_{i2,k}}{\sum_{j=1}^{N}})^2}$$

Then, the context distance \(d_{i, j}\) of the two data is as follows:

$$d_{i, j} = 0.089$$

There is a corresponding distance of each context attribute of any two data, then for \(i = 1, 2, \ldots, N\), there are N distances between any two data. But the importance of different context attributes is different and connected. So we combine single context dissimilarity and weight obtained above to get comprehensive context dissimilarity value.

3.3.2 Comprehensive Calculation of Context Dissimilarity

Context factors are often composed by a series of context attributes and the impact of a single context node on knowledge discovery is isolated. As shown above, context has characteristics of relevance and we should take account the combined effect of all context attributes so as to fully reveal the importance of context.

Through single context dissimilarity and the weights calculated by variance, we can get comprehensive distance of context after the weighted sum method. The context comprehensive distance takes both single context distances and the importance of different context attributes into consideration. The value represents the context context of two objects. The higher the value is, the greater the distance is, the smaller the similarity is.

There is a distance vector \((d_{x,y}, d_{x,y}^2, \ldots, d_{x,y}^N)\) between any two data. The component of the distance vector represents context distance value in different dimensions, which is single context dissimilarity. Based on the method of context attribute selection and calculation of corresponding
weight, the weighted sum of context distance in each dimension of two data can be used to reflect context distance of two data. Calculated as follows:

$$d_{x, y} = \sum_{j=1}^{N} w_j d_{x, y}^j$$

We can calculate the context distance of any two data through $d_{x, y}$ in order to select data based on the distance, modify results of basic classification algorithms and reuse knowledge, which can improve context-driven knowledge discovery.

4. CONTEXT-BASED KNOWLEDGE DISCOVERY

There is lack of effective guidance to traditional knowledge discovery at all stages to a certain extent, so that the results are unsatisfactory. Although rules obtained meet technical criteria, it is difficult to meet users’ needs. Adding context factors to the data mining process can get the decision-making closer to users and minimize the gap between data mining results and the reality.

Context-driven knowledge discovery focuses on adding context to the knowledge discovery process, so that the knowledge obtained is practical and instructive. The advantages are as follows compared with traditional knowledge discovery:

1. Guiding data preprocessing according to context, not only effectively reduce the amount of data, but also improve the efficiency of knowledge discovery.
2. Adding context in the data mining process can greatly improve the accuracy of the mining algorithms and compensate the defects of ignoring non-technical factors.
3. Context is viewed as one of the evaluation criteria in the selection process. On the basis of the traditional indicators such as novelty, interest and other technical factors, real context knowledge will be considered, so the knowledge obtained is closer to reality and more instructive at the same time.

The method about using context to guide data mining process is shown as below:

![Fig. 1 theoretical framework of context based knowledge discovery](image)

4.1 Steps of Context-Based Knowledge Discovery

Improvement of context reflects in various aspects of knowledge discovery. We can get more applicable rules through context to compensate for deficiencies of knowledge discovery. The basic steps of context based knowledge discovery are as follows:

1. Task analysis to determine knowledge discovery problem to be solved and identify context factors. First, searching for context knowledge base. Through context matching, find out whether there is existing knowledge that can be used directly. If there is, recommend the knowledge directly to users to solve the existing problems. If there is not, analyze context attributes and retain the largest several context attributes that affect the problem. Then find the database to calculate context distance of new problems with existing data. The smaller this distance is, the closer the two data is.
2. Except for preprocessing selected data to eliminate noise, outliers and vacancies of traditional knowledge discovery, context attributes needed to be processed in accordance with classification clean-up standards. In the process of data preprocessing, not only the original data but also the context data needs to be preserved in order to use context as guiding standards of knowledge discovery.
3. In order to get measures to deal with problems, we implement data mining algorithms on the basis of processed data. Compared with traditional algorithms, the algorithm is improved through the combination of context factors. It can not only retain the advantages of original data mining technology, but also make up for the deficiencies of existing technology. The improved algorithms reflect unique advantages of context and mining results can be more suitable to users’ needs.
4. After the implementation of mining algorithm, we can apply mining knowledge to real decision-making to solve the current problems. Traditional knowledge discovery ends, but context-driven knowledge discovery aims to better meet the spiral process of knowledge accumulation and knowledge reuse can be satisfied. Therefore, we need to make knowledge associated with current context in order to achieve knowledge store and reuse.

![Fig. 2 steps of context based knowledge discovery](image)

The following figure shows the theoretical framework and technologies of context based knowledge discovery. Context emphases and technologies in different stages are different. There will be a detailed analysis and introduction about key technologies in the following.
4.2 Data Selection Based on Context Similarity

At present, the explosive growth of data makes knowledge discovery inefficient; on the other hand, the relevance of data is difficult to guarantee, which will greatly affect the results of knowledge discovery. How to select data related to the task has been a key issue affecting the accuracy of knowledge discovery.

The data preprocessing of traditional knowledge discovery is designed to eliminate noise and redundancy, etc., which does not distinguish between data cleaning and data selection. After this stage, all complete data recorded will be used while modeling directly. However, for an enterprise, hundreds of millions of business data grow, so the idea of making use of all data is not realistic. Another problem is that many intelligent models require suitable data size. If the dataset is too large, analysis result is less than ideal. In addition, a large dataset leads to not only inefficient mining, but also the mixture of useful and useless rules. Thus, the selection of data related to the problem directly becomes a focus of the study.

Based on the methods above, while facing a new problem, we can represent data in the form of tuples. Then calculate the dissimilarity of the certain data and existing data in order to select the data with the most similar context while modeling. Based on this, we propose context distance-based data selection method. The method makes use of context as identification standard, which transforms data-driven into task-driven and makes the results more objective and effective.

![Flowchart of data selection based on context similarity](image)

Fig. 3 flowchart of data selection based on context similarity

Through computing context distance of test set and training set, the context-based data selection method extracts data with the most similar context attribute from the training set and implements algorithms on the dataset after context selection. On the other hand, the amount of data can be reduced after calculation of context distances.

4.3 Improved Knn Algorithm Based on Context

This section aims to study the combination of algorithms and context, which is the criteria of class label determination. Determining classification labels by the context similarity can make results more objective and avoid blind decision-making.

KNN stands for a kind of distance-based classification algorithms. First calculate the distance of the data to be classified and the training dataset. Then we get several “neighborhood” data points from the training set. Next assign the class label belonging to the majority of the data to be classified so as to determine the category. However, this algorithm focuses on statistically similarity and majority and does not consider other potential influencing factors, such as context.

KNN tends to assign the majority class of the nearest data points to the test set, leading to an important issue. When the amount of data is too large, the same value of classification attributes leads to inconsistent classes, which would affect the classification results. On the basis of selected data through context-based knowledge discovery, we can add context into distance-based classification algorithms, of which KNN algorithm is representative.

For test data $T_0$, the category of the nearest k data in the training set is $L_i (i = 1, ..., m)$ and the number of each category is $|L_1|, |L_2|, ..., |L_m|$ separately, which satisfies $|L_1| + |L_2| + ... + |L_m| = k$. $|L_i| (i = 1, ..., m)$ represents the number of data belonging to class $L_i$ among k data. During the k data, there exists context distance between each data belonging to class $L_i$ and the test data. Calculate the variance $Var_{Li}$ of the $|L_i|$ context distances.

The traditional method is to assign the majority class to the test data through the similarity of classification attributes. Different from previous methods, we assign the corresponding class label of data with the minimum variance to the test data after comparing the m context variances. The smaller the variance is, the smaller the interference is due to different context factors. If we assign the class label of the data with the minimum variance to the test data, the accuracy of classification can be improved.

Context-based KNN algorithm applies context distance to classic KNN algorithm. This method not only retains the idea of the nearest neighbors of traditional KNN algorithm, but also adds context standards to determine classification results, which is more targeted.

The process of context-based KNN is as follows:

**Input:** the number of k nearest neighbors; target object to be classified; dataset D

**Output:** class label of target object

**Methods:**

1. Calculate distances of natural attributes between target data and dataset;
2. Sort k-nearest neighbor in ascending order;
3. Calculate context distances of target object and the k data;
(4) calculate context variance belonging to each category among the k-nearest neighbors and assign the class label of the data with the smallest variance to the target data point.

5. **EMPIRICAL STUDY OF CONTEXT-BASED KNOWLEDGE DISCOVERY IN AUTO SALES**

Beijing Hyundai Motor Company's daily sales data documented customer purchasing situation. But on the one hand, sales staff do not know how to select data to extract rules; on the other hand, the effect of conclusion obtained from data analysis is not ideal in real car sales. How to develop more effective marketing strategies depending on existing datasets has become the focus of decision makers.

Our data comes from customers' information documented by Beijing Hyundai Motor Company from January to September, mainly containing cars' information, customers' information and information of car purchasing.

Traditional classification attributes are shown in Table 1 and 2, including cars' information and customers' information:

<table>
<thead>
<tr>
<th>Table 1 cars' information</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIN_NO</td>
</tr>
<tr>
<td>CAR_CD</td>
</tr>
<tr>
<td>CAR_NAME</td>
</tr>
<tr>
<td>MODEL_CD</td>
</tr>
<tr>
<td>OCN</td>
</tr>
<tr>
<td>MODEL_NAME</td>
</tr>
<tr>
<td>EX_COLOR_CD</td>
</tr>
<tr>
<td>EX_COLOR</td>
</tr>
<tr>
<td>IN_COLOR_CD</td>
</tr>
<tr>
<td>IN_COLOR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2 customers' information</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUST_NO</td>
</tr>
<tr>
<td>ID</td>
</tr>
<tr>
<td>CUSTOMER_NAME</td>
</tr>
<tr>
<td>TELEPHONE#1</td>
</tr>
<tr>
<td>TELEPHONE#2</td>
</tr>
<tr>
<td>BIRTH_YMD</td>
</tr>
<tr>
<td>SEX</td>
</tr>
<tr>
<td>JOB_CD</td>
</tr>
<tr>
<td>OCCUP_CD</td>
</tr>
<tr>
<td>INCOME_CD</td>
</tr>
<tr>
<td>SCHL_CD</td>
</tr>
<tr>
<td>HOBBY</td>
</tr>
<tr>
<td>PUR_MOTIVE</td>
</tr>
</tbody>
</table>

Compared with traditional algorithms, context based knowledge discovery improves by the combination of context factors. In this case, context information of sales data is shown in Table 3.

<table>
<thead>
<tr>
<th>Table 3 context information</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEALER_CODE</td>
</tr>
<tr>
<td>SALESMAN_NO</td>
</tr>
<tr>
<td>DEALER_NAME</td>
</tr>
<tr>
<td>SALE_DATE</td>
</tr>
<tr>
<td>ORD_TYPE</td>
</tr>
<tr>
<td>SALESMAN</td>
</tr>
<tr>
<td>POST_CD</td>
</tr>
<tr>
<td>ADDRESS</td>
</tr>
<tr>
<td>ADDRESS1</td>
</tr>
<tr>
<td>ADDRESS2</td>
</tr>
</tbody>
</table>

5.1 The Experiment Process

First, after data preprocessing including eliminating missing data and other noises, there are totally 63145 data used for analysis. Finally, determine 57169 data as training set and 5976 data as test set.

According to whether the value of cars' attribute "CAR_NAME" is Yilante or not, the overall data can be divided into two types "purshasing" and "non-purshasing". The positive class that stands for Yilante purchasing can be expressed as 1 and the negative class that stands for other car purchasing (not buying Yilante) can be expressed as 0. The ratio is 1:1, in accordance to equilibrium standards.

First remove classification attributes with values of more than 70% vacancies (eg, home phone, birth date, purchase purpose). Next conduct correlation test to remove redundant attributes of high correlation. Then transform data into available formats. Finally after feature selection, classification variables include consumers' demographic information such as age, sex, education, occupation and others, which can be seen in Table 4.

<table>
<thead>
<tr>
<th>Table 4 customers' classification attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
</tr>
<tr>
<td>AGE</td>
</tr>
<tr>
<td>JOB_CD</td>
</tr>
<tr>
<td>OCCUP_CD</td>
</tr>
<tr>
<td>INCOME_CD</td>
</tr>
<tr>
<td>SCHL_CD</td>
</tr>
<tr>
<td>HOBBY</td>
</tr>
<tr>
<td>PUR_MOTIVE</td>
</tr>
</tbody>
</table>

Traditional classification algorithm is based on customers' classification attributes above. The purpose is to mine relations of customers' attributes and car purchasing, so that we can predict whether the customer is a potential target or not.

After data preprocessing and the calculation of context variances, we get context attributes as shown in Table 5.

<table>
<thead>
<tr>
<th>Table 5 context attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SALE_DATE</td>
</tr>
<tr>
<td>ORD_TYPE</td>
</tr>
<tr>
<td>ADDRESS</td>
</tr>
</tbody>
</table>

Traditional classification algorithm is based on customers' classification attributes above. The purpose is to mine relations of customers' attributes and car purchasing, so that we can predict whether the customer is a potential target or not.
5.2 Experimental Results and Analysis

In order to verify context-based data selection, we can apply this method to reality through classic classification algorithms based on the original dataset and context-based dataset. The examining accuracies used for contrast are shown in Table 6.

<table>
<thead>
<tr>
<th>Accuracy classification algorithms</th>
<th>Original data</th>
<th>Selected data through context</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>59.90%</td>
<td>73.20%</td>
</tr>
<tr>
<td>Bayesian</td>
<td>76.60%</td>
<td>83.30%</td>
</tr>
<tr>
<td>C5.0</td>
<td>66.36%</td>
<td>67.64%</td>
</tr>
<tr>
<td>Neural network</td>
<td>77.93%</td>
<td>76.92%</td>
</tr>
<tr>
<td>CR Tree</td>
<td>75.92%</td>
<td>75.25%</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>77.42%</td>
<td>77.59%</td>
</tr>
</tbody>
</table>

At the same time, context-based KNN was applied to the original data and filtered data respectively. After horizontal comparison, the accuracy of improved KNN is from 66.70% to 76.90%, which validates the effectiveness of the context selection.

The data is real and due to data sources, the accuracy will be affected. However, based on the same data, the accuracy of context-based algorithm improved, which means the improvement effect is still quite significant.

Analyzing practically, “sale date” can influence traditional knowledge discovery results according to time differences; as previous example shows, purchasing tendencies from different regions are different and “purchase address” can eliminate the errors; corporate customers may need buses and taxi customers’ need is fixed, thus “order type” reflect the effect of this context factor.

6 CONCLUSION

In order to improve the practicability of mining results, this paper explores how to improve each stage of knowledge discovery to get actual rules so as to provide more effective decision-making criteria. This paper summarizes the deficiencies of the traditional knowledge discovery process, focusing on the framework and implementation methods of context-based knowledge discovery. First explore the context concept, composition and representation. Then propose probability distribution method to associate context with the data and sort the context attribute by calculating context variances, which can be used to filter context attributes. This paper designed context-based similarity assessment. On the one hand, in the data pre-processing stage, select similar data to the target object for data mining; on the other hand, make use of similar context as index and recommend associated knowledge to users for knowledge reuse.

We designed context-based KNN algorithm, introducing context into traditional KNN algorithm. The defects of previous algorithms that only consider classification attributes of data have been overcome. At the end of the article, through a case study of car purchasing data, we made use of different classification algorithms to mine selected data based on context and the accuracy improved compared to the original data. Meanwhile, the accuracy of context-based KNN algorithm has been raised than traditional algorithms.

ACKNOWLEDGEMENTS

The authors are very grateful to National Natural Science Foundation of China (Grant No. 71071151, 70921061) for research support.
REFERENCES


[2]. Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques[M]. Beijing: China Machine Press. 2007.


