An Inconsistency Problem in Data Discretion
Using Equal Width Interval Approach

Chien-Hsing Wu
Department of Information Management
National University of Kaohsiung
Taiwan
chwu@nuk.edu.tw

Hiroaki Ishii
Graduate School of Information Science
and Technology
Osaka University, Japan
ishii@ist.osaka-u.ac.jp

Abstract

Data discretion is one of the most important issues in data mining for continuous datasets. However, there is less attention that has been paid to the inconsistency of dataset that is discreted. In consequence, mining mechanism such as classification may face a difficulty in providing the classification outputs. We first reveal the problem of record inconsistency, and then describe a model that helps efficiently reveal the record inconsistency that do exist datasets that are discreted. The main outputs of the described model include revealed inconsistent records and consumed processing time. Fifteen continuous real-life datasets that are discreted by using the binning technique of S-plus histogram binning algorithm with equal width interval technique are tested. There are results obtained indicating that: (1) 38.89% of the discreted datasets contain inconsistent records and 22.22% of the discreted datasets have more than 20% amount of inconsistent records.

Keywords: Data mining, Discretion, Record consistency

Background

Data mining (DD) is one of the active research domains that is linked to data management, information representation, and machine learning techniques, in particular the volume of data generated increases rapidly [1-9]. If thoroughly review the DD research tasks, there are five major stages: data collection, the collected data preprocessing, pre-processed data mining, outputs collection, and output implementation and evaluation. The data collection basically is to gather real-life (sometimes artificial) data. Pre-processing deals mainly with data refinement and reconstruction of datasets, consistency of multi-typed datasets, elimination of redundant attributes, combination of highly correlative attributes, and
discretization of numeric attributes. The mining mechanism generally performs association, classification, regression, clustering, or summarization to explore knowledge that is significantly interesting, meaningful, and decision-supportable. Outputs contain the discovered knowledge that can be either documented as a report or used in intelligent systems to support in making decision. The work of implementation and evaluation is about the use of mined knowledge and test of the mining processes. It is dealt to ensure the DD pursuits that include reliability, efficiency, validity, simplicity, and/or generality. Obviously, in the DD research system, the results of each phase flow down to the next, and will have a great impact on the final results.

Discretion is the conversion of continuous attributes to categorical ones in order for mining mechanism to perform knowledge discovery. However, conflicting records may occur that have the same conditions, but different conclusion [10]. Although many studies have presented various techniques to improve the performance of discretion, the problem of record inconsistency in a dataset is still a serious issue that may consequently influence the reliability of the mind knowledge [2][3][8][11-18]. It should be noted that indisputably same conditions resulting in different conclusions in our real life is a common situation. However, with respect to an induction-based knowledge discovery mechanism, it should be defined that the same conditions must produce a single conclusion. For example, it will be a meaningless mined knowledge that if you study quite hard, you may or may not pass your final exam. Consequently, the investigation of record consistency for discreted datasets is quite important for data mining. A discreted dataset usually contains many attributes and many records. It is a highly labor-consumption task to investigate the conflicting records. We therefore adopt Structured Quesy Language (SQL) to develop a record consistency investigation model to help efficiently test discreted datasets [19-21]. The primary outputs of the described model include explored inconsistent records and consumed running time.

Record Inconsistency Problem

Before explaining the record inconsistency problem, we briefly describe the discretion. The discretion is to group numeric data for
each attribute in dataset. Figure 1 illustrated this in a graphic manner with equal width interval (EWI). It should be noted that there are many techniques used to group numeric data. However, this is beyond our research focus. The EWI has been often used as a conversion mechanism to generate nominal values from continuous ones. The EWI deals with the sorting the observed values of a continuous attribute and dividing the range of observed values for the variable into $k$ equally sized bins, where $k$ is a parameter predefined by the user. If a variable $x$ is observed to have values bounded by $x_{\text{max}}$ and $x_{\text{min}}$, then this method computes the equalized bin width as $(x_{\text{max}} - x_{\text{min}})/k$. As a result, the set of granules can be expressed as $G = \{G_1, G_2, \ldots, G_k\}$. The conversion function is defined as follows. Each nominal granule has the same continuous boundary, but the number of record may be different.

$$C_{\text{EWI}}(x_i) = \begin{cases} G_m, & \text{if } x_{\text{min}} + (m+1)d \leq x_i \leq x_{\text{min}} + md \\ G_k, & \text{if } x_i = x_{\text{max}} \\ \text{null} & \text{otherwise} \end{cases}$$

where

- $x_i$: the $i^{th}$ data.
- $G_m$: the $m^{th}$ granule that $y$ is grouped into, $m = 1, \ldots, k$, $k$: the number of granules.
- $x_{\text{min}}$: the minimum of the data.
- $x_{\text{max}}$: the maximum of the data.
- $d$: the equalized interval for $k$ granules.

![Figure 1: Information granularity](image-url)
It is very possible that inconsistent records may occur while performing the conversion operation. Conflicting records are very possible to produce dead-end leaf while constructing a decision tree [10]. In other words, conflicting records will stop returning decision rules. It is believed that conflicting records will very possibly result in two serious problems. One is that it is irrelevant to compute gained information for both attributes and classes if a discreted dataset contains too many inconsistent records. The other is that some information of the collected data sources will be lost if too many inconsistent records are eliminated.

To further deal with such a serious problem while discretizing the continuous attributes, an artificial inconsistency dataset is firstly generated as an example that has the record inconsistency problem to describe this concern in detail and secondly conduct an empirical investigation for real life datasets. The generated dataset contains 28 discreted records. Each record is assigned an ID. Also assume that the dataset has 7 attributes as well as a class with 7 labels and is ordered by the values of both attributes and class. The value domain of attributes for this dataset are: $V_A = \{A1, A2, A4, A5, A6, A7\}$, $V_B = \{B1, B2, B3, B5, B6, B7\}$, $V_C = \{C1, C2, C4, C5, C6, C7\}$, $V_D = \{D1, D2, D4, D5, D6, D7\}$, $V_E = \{E1, E2, E3, E4, E5, E6\}$, $V_F = \{F2, F3, F4, F5, F6\}$, $V_G = \{G1, G2, G3, G4, G5, G7\}$, and $V_{CSS} = \{CSS1, CSS2, CSS3, CSS4, CSS5, CSS6, CSS7\}$. If look at the records numbered from 1 to 7, it is found that the combinations of attributes are all equal to (A1, B2, C2, D4, E1, F2, G2), but have different conclusions that are (CSS1, CSS2, CSS3, CSS5).

Since inconsistent records are regarded as those that have the same conditions, but different conclusions, any record that has a single attribute values is considered to be consistent in a discreted dataset. Moreover, in order to further conveniently look at the inconsistency problem in depth and find a solution in general, four types of subdatasets based on the relations between attribute values and class values for a discreted dataset are defined as follows.

$D_{\text{initial}}$: A dataset that contains all records in a discreted dataset. For example, $D_{\text{initial}} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28\}$ in the artificial dataset. $D_{\text{single}}$: A dataset that contains every single
record of which the combination of attribute value is unique. For example, \( D_{\text{single}} = \{8, 9, 10, 18, 23, 27, 28\} \) in the artificial dataset. \( D_{\text{multiple}} \): A dataset that contains record sets. Each set has an identical combination of attribute value and the number of records is equal to or greater than 2. For example, \( D_{\text{multiple}} = \{\{1, 2, 3, 4, 5, 6, 7\}, \{11, 12, 13, 14\}, \{15, 16, 17\}, \{19, 20\}, \{21, 22\}, \{24, 25, 26\}\} \) in the artificial dataset. Note that \( D_{\text{multiple}} \) is equivalent to the difference set of \( D_{\text{initial}} \) and \( D_{\text{single}} \). \( D_{\text{SameConclusion}} \): A dataset that contains record sets. Each set is a member of \( D_{\text{multiple}} \) and has a unique conclusion. For example, \( D_{\text{SameClass}} = \{\{19, 20\}, \{24, 25, 26\}\} \) in the artificial dataset.

Apparently, the inconsistent records only exist in \( D_{\text{multiple}} \). Each element (a subset) in \( D_{\text{multiple}} \) needs to be further detected whether or not the records have the same conclusions. Any element of which the conclusions are constant is eliminated from \( D_{\text{multiple}} \) to derive the set that contains all inconsistent record. For example, \( S_{\text{SameClass}} \) containing two record sets of \{19, 20\} and \{24, 25, 26\} are such kind of elements in \( S_{\text{multiple}} \). The remainder then forms a subset that is denoted by \( D_{\text{DiffClass}} \) that contains the inconsistent record which are \{\{1, 2, 3, 4, 5, 6, 7\}, \{11, 12, 13, 14\}, \{15, 16, 17\}, \{21, 22\}\}. \( D_{\text{DiffClass}} \) is then used to investigate the initial dataset. However, for the algorithmic aspect, formula (1) is not just a simple algebra that can be used to explore inconsistent records. The size of dataset, number of attributes, number of values that each attribute can take on, and the number of values that the class can take on are all variables that may cause the problem very complex. In order for the solution to be generalized, the study needs to solve two problems. One is to separate \( D_{\text{single}} \) and \( D_{\text{multiple}} \) from \( D_{\text{original}} \), the other is to eliminate all subsets in \( D_{\text{SameClass}} \) from \( D_{\text{multiple}} \).

Proposed Model

The proposed model basically is developed by using a divide-group-join strategy. It contains three procedures: grouping, dividing, and joining. Grouping is to deal with the \( D_{\text{single}} \) and \( D_{\text{multiple}} \) separations from the initial dataset. The \( D_{\text{SameClass}} \) is derived by dividing while joining returns the final results. In Figure 2, there are three operations, \( OP_A \), \( OP_B \), and \( OP_C \) to complete the whole process. Each operation represents an SQL statement for a defined purpose. The
objective of each operation is listed in Table 1 where operations and their corresponding SQL statements with pseudo format are contained. By connecting all procedures, the defined SQL statements used to return inconsistent records is formed.

![Diagram of operations](image)

Figure 2: The operations of record consistency

<table>
<thead>
<tr>
<th>Operation</th>
<th>SQL Statements (pseudo format)</th>
<th>dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP_A</td>
<td>Select *, Count(dataset.a); From D_i; Group by all_attribute + class; Order by all_attribute + class; Into Table D_a</td>
<td>D_i</td>
<td>The initial discreted dataset</td>
</tr>
<tr>
<td>OP_B</td>
<td>Select *, Count(dataset.b); From D_i; Group by all_attribute; Order by all_attribute; Into Table D_b</td>
<td>D_a</td>
<td>The created temporary dataset that lists the number of records of which the conditions and conclusion are constant.</td>
</tr>
<tr>
<td>OP_C</td>
<td>Delete From D_b : Where cnt_b&lt;2</td>
<td>D_c</td>
<td>The created temporary dataset from D_a that lists all kinds of combinations of conditions and their number of occurrences.</td>
</tr>
<tr>
<td>OP_{diffClass}</td>
<td>Select * : From D_a , D_i ; Into Table S_{diffClass}; Where D_a.all_attribute = S_i.all_attribute</td>
<td>D_{diffClass}</td>
<td>The final subset that contains all inconsistent records.</td>
</tr>
</tbody>
</table>
A demonstrated example

An example as mentioned in the above section is used to demonstrate how the proposed model functions in a simpler way. $D_A$ was listed in Table 2 where the number of records with same conditions and conclusions were contained. In $D_A$, we realized that some records were consistent and some were not. However, there was only one record in $D_A$ if they were consistent, since all conditions and conclusions were taken into account in terms of grouping. Therefore, we regrouped via all conditions and counted for $D_A$. The output was stored in $D_B$ that was listed in Table 3. We then eliminated those records from $D_A$ that returned only one record to obtain $D_C$ listed in Table 4, because they represented consistent records in $D_A$. We finally performed inner join for $D_A$ and $D_C$ to obtain the final results $D_{DfC|ass}$. The $D_{DfC|ass}$ was listed in Table 5. All intermediate subsets as well as the final results were tested correct.
In order for the proposed model to be able to both investigate the record inconsistency and reveal record inconsistency problem, 18 continuous real-life datasets were tested. The discretization technique employed was the EWI that involves sorting the observed values of a continuous feature and dividing it into $k$ equally sized granules, where $k$ is the number of granules defined by users [2]. The binning algorithm, S-Plus Histogram Binning Algorithm (SHBA) introduced by Spector [22] was utilized to determine the number of granules. The characteristics of the experiment were that (1) the number of datasets used is 18, (2) missing data was eliminated if occurs, (3) binning technique utilized is SHBA, (4) discretization technique employed is EWI, (5) outputs of the experiment is inconsistent records and processing time, and (6) objective is the investigation of record
inconsistency. The SHBA is expressed as SHBA: 

$$k = \text{max}(1, \text{integer}(2 \times \log(\text{difference})))$$

where $k$ is the number of granules and difference is the number of different values that an attribute has.

In addition, the average of run time was the mean of twenty trials that were performed. The results of the experiment were listed in Table 7. It was found that of 18 datasets that were investigated, 7 were not recognized. This implied that 38.89% of the discreted datasets contained inconsistent records. More particularly, there was four datasets that contained inconsistent records more than 20%. The one that showed the biggest percentage of inconsistent records was 37.2373%. The experimental results also confirmed that the proposed model did not consume unacceptable processing time.

Table 7: Record consistency investigation via EWI with SHBA

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Size</th>
<th>Num. of attributes</th>
<th>Num. of class</th>
<th>Pct. of record inconsistency</th>
<th>Avg. of run time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC198</td>
<td>194</td>
<td>33</td>
<td>2</td>
<td>0.0000</td>
<td>0.3040</td>
</tr>
<tr>
<td>BC569</td>
<td>569</td>
<td>30</td>
<td>2</td>
<td>0.0000</td>
<td>0.4680</td>
</tr>
<tr>
<td>Bupa</td>
<td>345</td>
<td>6</td>
<td>2</td>
<td>28.9855</td>
<td>0.2220</td>
</tr>
<tr>
<td>Sonar</td>
<td>208</td>
<td>60</td>
<td>2</td>
<td>0.0000</td>
<td>0.5910</td>
</tr>
<tr>
<td>Satellite</td>
<td>2000</td>
<td>36</td>
<td>6</td>
<td>0.0000</td>
<td>1.9900</td>
</tr>
<tr>
<td>Pageblock</td>
<td>5473</td>
<td>10</td>
<td>5</td>
<td>37.2373</td>
<td>0.5010</td>
</tr>
<tr>
<td>Letter</td>
<td>16384</td>
<td>16</td>
<td>26</td>
<td>3.0945</td>
<td>6.0260</td>
</tr>
<tr>
<td>Pendigit</td>
<td>3498</td>
<td>16</td>
<td>10</td>
<td>0.0000</td>
<td>1.2430</td>
</tr>
<tr>
<td>BC699</td>
<td>699</td>
<td>9</td>
<td>2</td>
<td>0.5722</td>
<td>0.2530</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>9</td>
<td>7</td>
<td>19.1589</td>
<td>0.1480</td>
</tr>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
<td>4.0000</td>
<td>0.1220</td>
</tr>
<tr>
<td>Segmentation</td>
<td>210</td>
<td>16</td>
<td>7</td>
<td>0.0000</td>
<td>0.2320</td>
</tr>
<tr>
<td>Shuttle</td>
<td>14500</td>
<td>9</td>
<td>7</td>
<td>34.0828</td>
<td>2.2100</td>
</tr>
<tr>
<td>Synthetic</td>
<td>600</td>
<td>61</td>
<td>6</td>
<td>0.0000</td>
<td>0.8900</td>
</tr>
<tr>
<td>Vehicle</td>
<td>846</td>
<td>18</td>
<td>4</td>
<td>0.0000</td>
<td>0.4610</td>
</tr>
<tr>
<td>Vowel</td>
<td>900</td>
<td>10</td>
<td>11</td>
<td>0.0000</td>
<td>0.3560</td>
</tr>
<tr>
<td>Waveform</td>
<td>5000</td>
<td>21</td>
<td>3</td>
<td>0.0000</td>
<td>2.1490</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>3</td>
<td>0.0000</td>
<td>0.1880</td>
</tr>
</tbody>
</table>

**Conclusion**

This paper has briefly described the problem of record inconsistency and presented a model that is constructed by using SQL to help efficiently investigate the inconsistent records for discreted datasets. It has been demonstrated successful as a record inconsistency detector for any dataset
that is discreted. The model can be also directly used for discrete datasets. The equal width interval technique with SHBA embedded is employed in the proposed model. The results implied that record inconsistency investigation was truly an essential issue to the discreted datasets used in DM research. In our experiment, many parameters are considered. For example, the determination of $k$ is based on Spector [22]. Other techniques may have influence the final results, which is one of our future research focuses. Although that we believe that other discretion techniques utilized and binning regulations employed may greatly affect the percentage of inconsistent records in a discreted dataset, it is encouraged that datasets used in the DM research be investigated before moving to the next stages. Moreover, as we have mentioned, the decision of the size of granules for continuous attributes that need to be discreted is facing a problem of dilemma that suffers from unacceptable amount of conflicting records to be contained if too large and incomprehensible knowledge to be discovered if too small. Therefore, a mechanism that can carry out a near optimal solution for record consistency and reliability of a discreted rule also is advantageous to the further research focuses.

References

[7] U.M. Fayyad, P. Stolorz, Data Mining and KDD: Promise and


[16] M. Pazzani, An Iterative-Improvement Approach for The Discretization of Numeric Attributes in Bayesian Classifiers,


[19] J. Han, Y. Fu, K. Koperski, O. Zaiane, DMQL: A Data Mining Query Language for relational Databases, DMKD-96 (SIGMOD-96 Workshop on KDD), Montreal, Canada, 1996.

