Unsupervised Data Discretization of Mixed Data Types

Jee Vang

Outline

- Introduction
- Background
- Objective
- Experimental Design
- Results
- Future Work
Introduction

- Many algorithms in data mining, machine learning, and artificial intelligence operate only on either numeric or categorical data
- Datasets often contain mixed types
- Discretization is the process of transforming continuous variables to categorical variables
  - Few discretization algorithms address interdependence between variables in dataset with mixed type
  - Even fewer address such concerns in the absence of class label in the dataset

Background - Discretization

- Discretization
  - Static vs dynamic
  - Supervised vs unsupervised
  - Local vs global
  - Top-down vs bottom-up
  - Direct vs incremental
- Only one known discretization algorithm that addresses dataset with mixed data types, is unsupervised, and considers variable interdependencies
  - Based on principal component analysis (PCA) and frequent itemset mining (FIM), PCA+FIM
Background – The Dataset

- Dataset consists of 272 patients with drug abuse problems treated from November 1997 to March 2003; 60 patients removed due to inadequate follow-up; 3 patients removed due to unavailable demographics data; end up with 209 patients
- A total of 13 variables were monitored
  - Binary: system type, technical violation, race, gender
  - Continuous: arrest, drug test, employment, homeless shelter, mental hospitalization, physical hospitalization, incarceration, treatment, age

Objective

- Quantitatively compare the preservation of correlation in the categorical domain after discretization in the continuous domain
- Benchmark PCA+FIM with equal-width (EW) and equal-frequency (EF) approaches
- Measuring how much correlation is preserved will be accomplished by using Spearman and Kendall correlation tests
Experimental Designs

**Procedure**
- 1. measure the pair-wise correlations in the continuous domain
- 2. input data set into discretization algorithms
- 3. measure the pair-wise correlations in the categorical domain
- 4. use Spearman or Kendall ranked-based correlation tests to observe much correlation is preserved between correlations in continuous (step 1) and categorical domain (step 2)

Experimental Designs – Discretization Algorithms

**PCA+FIM (Java, BLAS/LAPACK)**
- 1. normalize and mean center data
- 2. compute correlation matrix
- 3. compute eigenvalues/eigenvectors of correlation matrix; keep set of eigenvectors whose eigenvalues account for 95% of the variance
- 4. project data into eigenspace
- 5. discretize variables in eigenspace by generating cutpoints
- 6. project cutpoints back to original representation space

**EW (Data PreProcessor)**
- K intervals of equal-widths are produced

**EF (Data PreProcessor)**
- K intervals with equal frequency of data points are produced
Experimental Designs – Pair-wise Correlation Measures

- Continuous pair
  - Pearson
  - Kendall
  - Spearman
- Categorical pair
  - Phi
  - Mutual information
- Continuous-binomial pair
  - Point biserial

Results - Cutpoints

- Objective is not primarily to judge qualitatively (i.e. how meaningful are the cutpoints)
- PCA+FIM and EF produce less cutpoints
- EW produces more cutpoints
### Results – Comparing Pearson correlation to phi and mutual information correlation

#### Pearson - Phi

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA+FIM</td>
<td>0.15</td>
<td>0.09</td>
</tr>
<tr>
<td>EW</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>EF</td>
<td>0.13</td>
<td>0.07</td>
</tr>
</tbody>
</table>

#### Pearson – Mutual Information

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA+FIM</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>EW</td>
<td>-0.10</td>
<td>-0.06</td>
</tr>
<tr>
<td>EF</td>
<td>0.11</td>
<td>0.08</td>
</tr>
</tbody>
</table>

### Results – Comparing Spearman correlation to phi and mutual information correlation

#### Spearman - Phi

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA+FIM</td>
<td>0.14</td>
<td>0.09</td>
</tr>
<tr>
<td>EW</td>
<td>0.46</td>
<td>0.33</td>
</tr>
<tr>
<td>EF</td>
<td>0.12</td>
<td>0.07</td>
</tr>
</tbody>
</table>

#### Spearman – Mutual Information

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA+FIM</td>
<td>0.15</td>
<td>0.11</td>
</tr>
<tr>
<td>EW</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td>EF</td>
<td>0.09</td>
<td>0.07</td>
</tr>
</tbody>
</table>
### Results – Comparing Kendall correlation to phi and mutual information correlation

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA+FIM</td>
<td>0.16</td>
<td>0.10</td>
</tr>
<tr>
<td>EW</td>
<td>-0.25</td>
<td>-0.17</td>
</tr>
<tr>
<td>EF</td>
<td>-0.01</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Spearman</th>
<th>Kendall</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA+FIM</td>
<td>0.19</td>
<td>0.14</td>
</tr>
<tr>
<td>EW</td>
<td>-0.35</td>
<td>-0.20</td>
</tr>
<tr>
<td>EF</td>
<td>0.03</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### Results – Interpretation of correlation preservation

- If Pearson correlation is used to measure correlation in the continuous domain, PCA+FIM will produce a discretized dataset preserving the most correlation.
- If Spearman correlation is used to measure correlation in the continuous domain, EW will produce a discretized dataset preserving the most correlation.
- EF seems to preserve the least correlations in the categorical domain from the continuous domain.
- PCA+FIM shows consistency in correlation preservation.
Future Work

- Implement k-nearest neighbor approach in PCA+FIM discretization algorithm
- Test on other datasets

References