Point-Distribution Algorithm for Mining Vector-Item Patterns

Anne M. Denton, Jianfei Wu
Department of Computer Science
North Dakota State University, Fargo, ND

Dietmar H. Dorr
Research and Development
Thomson Reuters, St. Paul, MN
Motivation: Increasing Diversity of Data

- Vector data
- Groups of continuous data from multiple sources
- Could be result of feature extraction
- Item data
  - Binary with presence less frequent than absence
  - Could be item sets

<table>
<thead>
<tr>
<th>Obj. ID</th>
<th>Item Data (Boolean)</th>
<th>Continuous Data (Source 1)</th>
<th>Continuous Data (Source 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#001</td>
<td>0</td>
<td>0.3</td>
<td>0 0.2</td>
</tr>
<tr>
<td>#002</td>
<td>1</td>
<td>-2.1</td>
<td>0.1 0</td>
</tr>
<tr>
<td>#003</td>
<td>1</td>
<td>3.6</td>
<td>0 0</td>
</tr>
<tr>
<td>#004</td>
<td>0</td>
<td>0.8</td>
<td>0 0.1</td>
</tr>
<tr>
<td>#005</td>
<td>0</td>
<td>-1.4</td>
<td>0.2 0.3</td>
</tr>
<tr>
<td>#006</td>
<td>0</td>
<td>8.2</td>
<td>0.5 0.1</td>
</tr>
</tbody>
</table>

Vector Attributes
“Patterns of Usefulness”

• Supervised learning as well as some pattern mining approaches assume relationships exist.

• Find which vector attribute is most relevant to which item: Multi-dimensional feature selection.

• Find item sets that result in the clearest patterns (design of coatings).

• Establish relationship: Multi-dimensional hypothesis testing.
Problem Statement

• Identify items for which distribution of points with item differs significantly from overall distribution (figure 1-d)

• Related work: Is classification significant?

Point distributions

- All data points
- Points with item 1 (strong pattern)
- Points with item 2 (no strong pattern)
Common Approach in Bioinformatics

- Cluster, then look for enrichment of clusters
- Can miss significant relationships
Approach

- Define density, using kernel function (uniform kernel)
- Compare densities of points with item to densities of all points
- Previous approach used histograms
- Kullback-Leibler divergence quantifies difference between distributions directly

\[
D_{KL}^{(d)}(P||Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} \, dx_1, \ldots, dx_d
\]

\[
D_{KL}^{(d)}(P||Q) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x^{(i)})}{q(x^{(i)})}
\]
Algorithm (simplified)

- One parameter: Similarity threshold \( thresh \) (next slide)
- For each item
  - For all points with item
    - Find number of neighbors with item closer than \( thresh \) and divide by overall support of item: \( p(x) \)
    - Find total number of neighbors closer than \( thresh \) and divide by total number of points: \( q(x) \)
  - Calculate Kullback-Leibler divergence and compare with distribution of K-L divergences for random data
Parameter Choice

• Only parameter: Expected number of neighbors
• Choice of one intuitive (can be justified mathematically)
• Confirmed by experiments
Calculation of \textit{thresh}

- Calculate surface of cap of hypersphere

\[ S_{d-1}^{(\text{cap})} = \int_{0}^{\theta_t} S_{d-2} \ r^{d-2} \ d\theta = \int (d-1) \ C_{d-1} \ r^{d-2} \ d\theta \]

- Gives expression of support as function of \textit{thresh} that can be used for lookup

\[ \frac{1}{\text{support}} = \frac{\int_{0}^{\sqrt{1-\text{thresh}^2}} \frac{r^{d-2}}{\sqrt{1-r^2}} \ dr}{2 \int_{0}^{1} \frac{r^{d-2}}{\sqrt{1-r^2}} \ dr} \]
Genomics Application

- Finding protein domains that are related to a set of experiments in yeast
- Protein domains
  - Binary attributes
  - From Interpro database
- Gene expression data
  - All come from cell cycle experiments and are expected to represent related information
  - Four time series, each one consisting of 14 - 24 experiments
Results for Gene Expression Data

- Significance not known independently
- Results should be consistent over comparable experiments
  - Top right: Overlap
  - Bottom left: Significance that results are related

Table 1: Results for Gene Expression Data

<table>
<thead>
<tr>
<th></th>
<th>All</th>
<th>Alpha</th>
<th>Cdc15</th>
<th>Cdc28</th>
<th>Elu</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>259</td>
<td>114</td>
<td>117</td>
<td>160</td>
<td>166</td>
</tr>
<tr>
<td></td>
<td>119</td>
<td>67</td>
<td>85</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>134</td>
<td>86</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>7E-15</td>
<td>3E-12</td>
<td></td>
<td>173</td>
<td>107</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>2E-16</td>
<td>7E-12</td>
<td></td>
<td>198</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1E-7</td>
<td>0.027</td>
<td>5E-8</td>
<td></td>
</tr>
</tbody>
</table>
Labeled Data From Time Series

- Construct labeled data from time series sub-sequences
- Item data: membership in time series
- Noisy data (intentionally chosen)
- Allows varying item support by adding random walk data

```
<table>
<thead>
<tr>
<th>Vector Data</th>
<th>Time Series Items</th>
<th>Randomized Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-series 1</td>
<td>1 0 1 -4</td>
<td>1 0</td>
</tr>
<tr>
<td></td>
<td>-4 1 0 1</td>
<td>1 0</td>
</tr>
<tr>
<td></td>
<td>... ... ... ...</td>
<td>... ...</td>
</tr>
<tr>
<td>Time-series 2</td>
<td>0 3 0 0</td>
<td>0 1</td>
</tr>
<tr>
<td></td>
<td>3 0 0 -3</td>
<td>0 1</td>
</tr>
<tr>
<td></td>
<td>... ... ... ...</td>
<td>... ...</td>
</tr>
<tr>
<td>Random Walk Time-series</td>
<td>3 -2 1 -2</td>
<td>0 0</td>
</tr>
<tr>
<td></td>
<td>1 -2 1 2</td>
<td>0 0</td>
</tr>
</tbody>
</table>
```
Comparison Approaches

- Histogram-based approach from
  - A.M. Denton and J. Wu, KAIS, 2009
  - Summarizes density distributions as histograms

- Classification-based approach
  - Predict each item using classification (tree-based classifier in MATLAB)
  - Make prediction using 2-fold cross-validations
  - Calculate significance based on confusion matrix
Accuracy Depending on Item Support

- Clearly superior to comparison algorithms
- For very large item support classification may become competitive
Accuracy Depending on Amount of Noise

- Accuracy superior for all settings
- Not degrade much with added noise

![Graph showing accuracy depending on amount of noise.](image-url)
Performance

- Speed comparable to other algorithms
- Scaling with item support poorer
- Accuracy main motivation
- Algorithm most important for small item support
- Single parameter (expected # of neighbors) set to one
- Experiments confirm choice
Conclusions

• Solves an important problem

• Finding significant relationships between vectors and items or item sets

• Use of Kullback-Leibler divergence better justified theoretically than histograms

• Application to genomic data gives consistent results

• Accuracy much improved on semi-artificial data (constructed from real time series)