Multiple Instance Ranking
http://reccr.chem.rpi.edu/MIRank

Charles Bergeron       Jed Zaretzki
Kristin P. Bennett    Curt Breneman
Mathematical Sciences  Chemistry
Rensselaer Polytechnic Institute
Troy, New York, United States of America
Multiple Instance Ranking

Summary

• This presentation introduces a framework that tackles a novel machine learning question arising from an investigation into an important chemistry problem.
• Multiple Instance Ranking (MIRank) is defined and formulated.
• A first working algorithm produces excellent results on several real and synthetic problems.
MULTIPLE INSTANCE RANKING

Outline

• Motivation
  – from a Chemistry perspective
  – from a Machine Learning perspective

• Formulation
  – definition of Multiple Instance Ranking
  – algorithm as a bilinear optimization problem

• Experiments
  – datasets and experimental design
  – results and conclusions

• Outlook
  – from a Chemistry perspective
  – from a Machine Learning perspective
MULTIPLE INSTANCE RANKING

Outline

• Motivation
  – from a Chemistry perspective
  – from a Machine Learning perspective

• Formulation
  – definition of Multiple Instance Ranking
  – algorithm as a bilinear optimization problem

• Experiments
  – datasets and experimental design
  – results and conclusions

• Outlook
  – from a Chemistry perspective
  – from a Machine Learning perspective
Oesophagus

- Bioavailability (the ability of a drug administered orally to reach the bloodstream) is an important consideration to the pharmaceutical industry.
Stomach

- Pill is broken down into basic components.

Clozaril pill

Clozapine molecule

Motivation
Formulation
Experiments
Outlook

MULTIPLE INSTANCE RANKING

ICML, 8 July 2008
© Charles Bergeron et al., 2008. All rights reserved.
Liver & small intestine

- Successful drug compounds cross the hepatic and intestinal lining and make it to the bloodstream without being degraded, so that their medicinal effect may be felt.
- The rate limiting step in the metabolism of drugs by enzyme cytochrome CYP3A4 is hydrogen atom abstraction (removal).
Chemistry research goal

- To better understand the process of drug metabolism.
- To obtain knowledge about the site of metabolism where drug molecules undergo hydrogen abstraction under the effect of cytochrome CYP3A4.
Available data

- Database consists of small drug-like molecules (stick diagram).
- Features are computed for each hydrogen atom (small spheres).
- Hydrogens are grouped into sites of metabolism (different colors).
- For each molecule, the preferred site of metabolism is known.
- It is not known which hydrogen actually gets abstracted.
# Multiple Instance Ranking

## Chemistry to machine learning

### Problem statement
- **Computational chemistry**
  - For each molecule, find the site of metabolism
- **Machine learning**
  - For each box, find the preferred bag

### Motivation
- **Formulation**
- **Experiments**
- **Outlook**

<table>
<thead>
<tr>
<th>Level</th>
<th>Computational chemistry</th>
<th>Machine learning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Problem statement</strong></td>
<td>For each molecule, find the site of metabolism</td>
<td>For each box, find the preferred bag</td>
</tr>
<tr>
<td><strong>Top-level</strong></td>
<td>Molecules</td>
<td>Boxes</td>
</tr>
<tr>
<td><strong>Middle-level</strong></td>
<td>Sites of metabolism</td>
<td>Bags</td>
</tr>
<tr>
<td><strong>Bottom-level</strong></td>
<td>Hydrogen atoms</td>
<td>Items</td>
</tr>
</tbody>
</table>

### Diagram
- Molecules and sites of metabolism with attached numbers, indicating the degrees of preference.
- Boxes with stars indicating the preferred bags.
For each box (red rectangle), predict the preferred bag (green ellipse). Items are represented as parallelograms and stars, and the other bags are blue ellipses.
The need for MIRank

<table>
<thead>
<tr>
<th>Dataset particularities</th>
<th>Machine learning consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Descriptors are known for each item.</td>
<td>This is a multiple instance problem.</td>
</tr>
<tr>
<td>• Each box has exactly one preferred bag.</td>
<td></td>
</tr>
<tr>
<td>• An ambiguity exists as to which item in that bag determines the preference.</td>
<td></td>
</tr>
<tr>
<td>• It is not known how other bags within a box rank with respect to each other.</td>
<td>This is a partial ranking problem within each box.</td>
</tr>
<tr>
<td>• It is not known how bags compare against each other across boxes.</td>
<td></td>
</tr>
<tr>
<td>• Boxes may be very different from each other.</td>
<td>This is a hard problem.</td>
</tr>
</tbody>
</table>
Machine learning research goal

- To define the Multiple Instance Ranking (MIRank) setting.
- To develop a first working algorithm to solve MIRank problems.
- To prove the concept of MIRank on the CYP3A4 substrate and other real and synthetic datasets.
Outline

• Motivation
  – from a Chemistry perspective
  – from a Machine Learning perspective

• Formulation
  – definition of Multiple Instance Ranking
  – algorithm as a bilinear optimization problem

• Experiments
  – datasets and experimental design
  – results and conclusions

• Outlook
  – from a Chemistry perspective
  – from a Machine Learning perspective
Starting point: MIC

- Multiple Instance Classification (MIC) classifies bags based on their item descriptors.
- An active bag contains at least one active item.
- An inactive bag contains exclusively inactive items.
- There exists the ambiguity of what item in the active bag is active.
MULTIPLE INSTANCE RANKING

Motivation Formulation Experiments Outlook

MIC schematic

Items (stars and parallelograms) in active bags (green ellipses) and inactive bags (blue ellipses) are shown.
At least one item (star) in each active bag (green ellipse) is above the decision curve (orange line).
MULTIPLE INSTANCE RANKING

Motivation    Formulation   Experiments   Outlook

MIRank Schematic

For each box (red rectangle), predict the preferred bag (green ellipse). Items are represented as parallelograms and stars, and the other bags are blue ellipses.
For each box (red rectangle), the ranking function (orange arrow) ranks highest the preferred bag (green ellipse).
For each box (red rectangle), the ranking function (sliding orange line) ranks highest the preferred bag (green ellipse).
MULTIPLE INSTANCE RANKING

Motivation    Formulation    Experiments    Outlook

Mathematical formulation

\[
\max_{i \in I} f(x_i) > \max_{j \in J} f(x_j)
\]

\[
\max_{i \in I} f(x_i) > f(x_j) \quad \forall \quad j \in J
\]

\[
f(X_i^T v_{i,J}) > f(x_j) \quad \forall \quad j \in J
\]

\[
f(x) = x^T w
\]

\[
v_{i,J}^T X_I w > x_j^T w
\]

\[
v_{i,J}^T X_I w - x_j^T w \geq 1 - \xi_{i,j}
\]

- Bag I is greater than bag J
- RHS max replaced with all items
- LHS max replaced with convex combination of all items in that bag
- Assume a linear model
- Apply model
- Allow for error by introducing empirical risk and margin
Bilinear optimization problem

\[ \min_{\xi, w, v_{I, J}} v e^T \xi + \| w \|_1 \]

subject to

\[ v_{I, J}^T X_I w - x_j^T w \geq 1 - \xi_{I, J} \quad \forall \quad (I, J, j) \]

\[ e^T v_{I, J} = 1 \quad \forall \quad (I, J) \]

\[ v_{I, J} \geq 0 \quad \forall \quad (I, J) \]

Empirical risk terms are nonnegative.

Empirical risk terms are nonnegative.
Modeling choices

- A linear model is chosen, because chemists are interested in easily interpretable models.
- The 1-norm is used to regularize, resulting in sparse models, again to facilitate chemical interpretation.
- The 1-norm is also used to penalize empirical risk, as is done in SVM. Also, this keeps the objective linear.
- The Mangasarian & Wild (2008) strategy of using the convex combination as a slick way of handling the uncertainty as to which item determines the preference is chosen.
Previous choices

- Support vector machine generalizations exist using integer variables (Andrews et al., 2003) and convex combinations (Mangasarian and Wild, 2008) to handle the ambiguity
Algorithm outline

- Solve the linear program for $\xi$ and $w$ while keeping the $v$’s fixed.
- Solve the linear program for $\xi$ and the $v$’s while keeping $w$ fixed.
- Repeat.

$$\begin{align*}
\min_{\xi, w, v_I, j} & \quad \nu e^T \xi + \|w\|_1 \\
\text{subject to} & \\
& v_I, J^T X_I w - x_j^T w \geq 1 - \xi_{I, j} \quad \forall \ (I, J, j) \\
& e^T v_I, J = 1 \quad \forall \ (I, J) \\
& v_I, J \geq 0 \quad \forall \ (I, J) \\
& \xi \geq 0.
\end{align*}$$
MULTIPLE INSTANCE RANKING

Outline

• Motivation
  – from a Chemistry perspective
  – from a Machine Learning perspective

• Formulation
  – definition of Multiple Instance Ranking
  – algorithm as a bilinear optimization problem

• Experiments
  – datasets and experimental design
  – results and conclusions

• Outlook
  – from a Chemistry perspective
  – from a Machine Learning perspective
MULTIPLE INSTANCE RANKING

Motivation   Formulation   Experiments   Outlook

CYP3A4 substrate dataset

- 227 molecules (boxes)
- 2272 sites of metabolism (bags)
- 4847 hydrogen atoms (items)
- 32 features per item
  - charge
  - hydrogen surface area
  - nonhydrogen surface area
  - hydrophobic moment
  - span
  - topological neighborhood properties
- Output information consists of one preferred bag per box
Experimental design

- Randomly split the dataset into training, validation and testing subsets consisting of 60%, 20% and 20% of boxes, respectively.
- Train MIC and MIRank on the training subset for 19 values of the tradeoff parameter $\nu$.
- The validation subset is used to select the best $\nu$.
- Results are recorded over the testing subset.
- This process is repeated 32 times.
Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MIC</th>
<th>MIRank</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CYP3A4 substrate</td>
<td>67.2% ± 6.6</td>
<td>70.8% ± 6.4</td>
<td>4.10·10^{-3}</td>
</tr>
</tbody>
</table>

- MIC algorithm is that of Mangasarian and Wild (2008).
- Metric is the percentage accuracy in predicting the preferred bag for each box, allowing for 2 guesses per box.
- Results are presented as a mean and std over the 32 runs.
- MIRank statistically outperforms MIC at a 5% significance level.
- MIRank improves the model by over 5%.
Further results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MIC</th>
<th>MIRank</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic-1</td>
<td>90.8 % ± 8.6</td>
<td>99.8 % ± 0.53</td>
<td>1.62·10^{-6}</td>
</tr>
<tr>
<td>Synthetic-2</td>
<td>96.8 % ± 4.6</td>
<td>99.1 % ± 1.8</td>
<td>1.31·10^{-2}</td>
</tr>
<tr>
<td>Synthetic-3</td>
<td>95.5 % ± 8.3</td>
<td>99.9 % ± 0.38</td>
<td>5.84·10^{-3}</td>
</tr>
<tr>
<td>Synthetic-4</td>
<td>95.7 % ± 5.2</td>
<td>99.7 % ± 0.91</td>
<td>1.46·10^{-4}</td>
</tr>
<tr>
<td>Census-16h</td>
<td>52.8 % ± 17.4</td>
<td>60.3 % ± 15.1</td>
<td>4.51·10^{-2}</td>
</tr>
<tr>
<td>Census-16l</td>
<td>46.2 % ± 17.7</td>
<td>57.5 % ± 16.0</td>
<td>3.92·10^{-4}</td>
</tr>
</tbody>
</table>

- For all datasets, MIRank statistically outperforms MIRank.
- For the census datasets, MIRank improves the model by 14-24%.
- The paper goes into greater depth about these datasets.
Discussion

• Problems fitting into the Multiple Instance Ranking paradigm are better solved using MIRank models than other methods.

• Forcing MIRank problems into a MIC paradigm is not successful.
Software and data

http://reccr.chem.rpi.edu/MIRank

- We are making our CYP3A4 substrate data available.
- We are also making our MIRank algorithm Matlab source codes available, as well as our implementation of the MIC algorithm (Mangasarian & Wild, 2008).
- Look out for it online!
- Contact me for further information.
MULTIPLE INSTANCE RANKING

Outline

• Motivation
  – from a Chemistry perspective
  – from a Machine Learning perspective

• Formulation
  – definition of Multiple Instance Ranking
  – algorithm as a bilinear optimization problem

• Experiments
  – datasets and experimental design
  – results and conclusions

• Outlook
  – from a Chemistry perspective
  – from a Machine Learning perspective
Chemistry extensions

- Chemical interpretation of results is a paper of its own.
- Increase the number of molecules in CYP3A4 database.
- Build databases and models for new substrates, such as CYP2D6 and CYP2C9.
- Develop novel descriptors believed to be indicative of hydrogen abstraction.
Machine learning extensions

- Implementation of nonlinear model using kernels (already formulated in the paper) to compare results with linear model.
- Compare model choice of 1-norm with other possibilities for empirical risk and regularization terms.
- Adapt recent large scale SVM algorithms to make MIRank more scalable and efficient at finding local minima.
- Use integer programming or cutting plane algorithms to find global minima (at much greater computational cost).
Future applications

- For each country, predict the city that contains the most profitable coffee shop.
- For a given state/province/länder/département, predict the electoral district that contains the most effective politician (the one that delivers the most subsidies to his constituents).
- For a given molecular class, predict the molecule that contains the conformation with the highest efficacy in inhibiting the human immunodeficiency virus (HIV).
- For each Olympic event, predict the nationality of the winning athlete.
- For each document, find the paragraph/passage that contains the most relevant phrase/sentence/word.
Summary

- This presentation introduces a framework that tackles a novel machine learning question arising from an important chemistry problem.
- Multiple Instance Ranking (MIRank) is defined and formulated.
- A first working algorithm produces excellent results on several real and synthetic problems.
Multiple Instance Ranking
http://reccr.chem.rpi.edu/MIRank

Charles Bergeron
Kristin P. Bennett
Mathematical Sciences
Rensselaer Polytechnic Institute
Troy, New York, United States of America

Jed Zaretzki
Curt Breneman
Chemistry