Application of Classification Association Rule Mining for Mammalian Mesenchymal Stem Cell Differentiation

Speaker: Yanbo J. Wang
Information Management Center, China Minsheng Banking Corp., Ltd.

Co-work with: Weiqi Wang, René Bañares-Alcántara, Zhanfeng Cui
Department of Engineering Science, University of Oxford
Frans Coenen
Department of Computer Science, University of Liverpool
BACKGROUND
Bone Marrow Stem Cells

In our study, we focus on MSC investigation
Our study is mainly concentrated on MSC differentiation.
Mesenchymal Stem Cell Differentiation

Differentiation Potential (Fate) of MSCs

- Osteogenesis
- Tendonogenesis
- Chondrogenesis
- Myogenesis
- Adipogenesis
- Neurogenesis
MOTIVATION
We are interested in finding the way how MSC to be differentiated.

The scattered data on this (MSC) study is available online – we can extract and collect such data from online research/academic repositories, e.g. MEDLINE.

In general, such differentiation problem can be simply modelled as a Classification problem in Data Mining.

Our study is concerned with the *single-label* Classification task – assigning exactly one predefined class (differentiation fate) to each “unseen” (MSC) data record.

There are many Classification approaches/mechanisms available, i.e. Artificial Neural Network, Support Vector Machine, Naive Bayes, K-Nearest Neighbour, Classification Association Rule Mining, etc.
CLASSIFICATION ASSOCIATION RULE MINING
In our study, we select to use the Classification Association Rule Mining (CARM) approach.

CARM offers the following advantages:

- The approach is efficient during both the training and categorisation phases, especially when handling a large volume of data.
- The classifier built in this approach can be read, understood and modified by humans, whereas other classifiers cannot.
- CARM is relatively insensitive to noise data.
- In previous studies, CARM was reported to offer good classification accuracy.

CARM strategically solves the traditional Classification problem by applying Association Rule Mining (ARM) techniques.
CARM aims to extract a set of Classification Association Rules (CARs) from a class-transactional database $D_{C-T}$. Let $D_T$ be a (binary-valued) transactional database, and $C = \{c_1, c_2, \ldots, c_{|C|-1}, c_{|C|}\}$ be a set of predefined class labels, $D_{C-T}$ is described by $D_T \times C$.

A CAR describes an implicative co-occurring relationship between a set of binary-valued data attributes and a predefined class, expressed in the form of an “$X \Rightarrow c_i$” rule, where $X$ is an itemset found in $D_T$ and $c_i$ is a predefined class in $C$.

A CAR is said to be valid when the support of $X$ and $c_i$ exceeds a user supplied $\text{minsupp}$ (support threshold), and the confidence of this CAR exceeds a user supplied $\text{minconf}$ (confidence threshold).

- support ($X \cup c_i$) = count ($X \cup c_i$ in $D_{C-T}$) / $|D_{C-T}|$.
- confidence ($X \Rightarrow c_i$) = support ($X \cup c_i$) / support ($X$).
PROCEDURES
A domain-dependent database containing total 375 parameters that are believed to influence the MSC differentiation has been built and can be accessed online at “http://www.oxford-tissue-engineering.org/forum/plugin.php?identifier=publish&module=publish”.

Each record in the MSC database that describes a (real-life) experiment of MSC differentiation, was read, extracted and collected from such research/academic papers (note that paper reference is also recorded in the database).

The size of this database was 203 records (as reported in the conference paper), and now it has been increased up to 501 records.

In this database, the key (most significant) parameters include: in vivo/vitro, culture medium, supplement, monolayer/3D culture, substrate/scaffold, cell seeding density, result, etc.
<table>
<thead>
<tr>
<th>record_no</th>
<th>title</th>
<th>authors</th>
<th>institute</th>
<th>species</th>
<th>tissue</th>
<th>phenotype</th>
<th>condition</th>
<th>supplement</th>
<th>control</th>
<th>type</th>
<th>differentiation</th>
<th>added</th>
<th>update</th>
</tr>
</thead>
<tbody>
<tr>
<td>00006</td>
<td>m,v</td>
<td>yes</td>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12345</td>
<td>m,v</td>
<td>yes</td>
<td>2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>67890</td>
<td>m,v</td>
<td>yes</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11111</td>
<td>m,v</td>
<td>yes</td>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22222</td>
<td>m,v</td>
<td>yes</td>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33333</td>
<td>m,v</td>
<td>yes</td>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44444</td>
<td>m,v</td>
<td>yes</td>
<td>2009</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55555</td>
<td>m,v</td>
<td>yes</td>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>66666</td>
<td>m,v</td>
<td>yes</td>
<td>2011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>77777</td>
<td>m,v</td>
<td>yes</td>
<td>2012</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>88888</td>
<td>m,v</td>
<td>yes</td>
<td>2013</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99999</td>
<td>m,v</td>
<td>yes</td>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

All times are GMT, the time now is 2-1-2006 22:52
Data Processing

Original data in MSC database

Schema

Input data

CARM

Normalised data

De-noise

Generated rules
RESULTS
Experiments were run on a 2.00 GHz Intel(R) Core(TM)2 CUP with 2.00 GB of RAM running under Windows Command Processor.

The evaluation was performed using the CMAR (Classification based on Multiple Association Rules) algorithm although any other CARM classifier generator (i.e. CBA, CPAR, TFPC, etc.) could equally well have been used. The CMAR software can be downloaded from “http://www.csc.liv.ac.uk/~frans/KDD/Software/CMAR/cmar.html”.

The evaluation undertaken used a support threshold value \((\text{minsupp})\) of 1% and a confidence threshold value \((\text{minconf})\) of 50%.

The evaluation was performed with the Ten-fold Cross Validation (TCV) accuracy setting.

The classification (prediction) accuracy was 77.04% with 203 data records (as reported in the conference paper).

The up-to-dated classification accuracy is 90.4% with 501 data records.
Interesting Rules

Rules that are generated by the CMAR software
With regard to a particular fold in the TCV process, there were 163 CMAR rules generated from the input data, which is around 182 data records (203 records × 90% of the database).

**Rule # 49:** \{in vitro + monolayer + human donor + DMEM + TGFβ1 + plastic substrate\} $\Rightarrow$ \{chondrogenesis\} [100.0%], which can be interpreted as: in monolayer culture in vitro, human MSCs are most likely to undergo chondrogenesis in the presence of cell culture medium DMEM (Dulbecco’s Modified Eagle’s Medium) and growth factor TGFβ1 (Transforming Growth Factor β1), on plastic substrate.

**Rule # 86:** \{DMEM + FBS + ascorbate-2-phosphate + Dex\} $\Rightarrow$ \{osteogenesis\} [93.33%], which can be interpreted as: in DMEM medium supplemented with FBS (Fetal Bovine Serum), MSCs are very likely to be induced to osteogenesis under the stimuli of ascorbate-2-phosphate and Dex (Dexamethasone) together.
Based on the total 501 data records, the new training dataset consists of around 450 data records (501 records \( \times \) 90\% of the database). There were 295 CMAR rules generated, among which many are found to be interesting.

**Interesting rules**

**Rule # 27:** \{pyruvate + proline\} \( \Rightarrow \) \{chondro\} \[100.0\%\]
- Pyruvate: important in metabolic pathways, may potentially help promote MSC chondrogenesis.
- Proline: a catalyst in biochemical reactions, may facilitate chondrogenesis.
- Not 100\% sure, but we just did not realise them yet!

**Rule # 188:** \{transferrin + selenous acid + dexamethasone\} \( \Rightarrow \) \{chondro\} \[91.17\%\]
- Transferrin: participate in the immune system, prevent bacteria from survival, not contrary for transferrin to support chondrogenesis.
- Selenous acid: **highly toxic, usually fatal**, needs further investigation.
CONCLUSION & FUTURE WORK
Conclusion and Future Work

- CARM is a promising method to discover rules involved in MSC differentiation.
- The classification accuracy of this approach is good.
- Some rules have been found to be interesting, and need further investigation.

- In the future, we would like to continuously increase/expand the size of the MSC database.
- For this (MSC differentiation) study, we also want to compare the performance among various CARM algorithms, i.e. CMAR, CBA, CPAR, TFPC, etc.
We are the most grateful if you can share your MSC data online!


THANK YOU!