Fingerprint Clustering with Missing Values

Figueroa et al CATS 2005, Bonizzoni et al CPM 2006

Algorithmics Group, Dept. of Computer Science, University of Liverpool

Speaker: Cindy Li
Supervised by: Leszek Gąsieniec & Prudence Wong
What is CMV(p) problem?

- **INSTANCE:** A set $F$ of $n$ 0-1-N vectors with at most $p$ Ns per vector. Each vector has length $l$.
- **SOLUTION:** A partition of $F$ into disjoint subsets $F_1, F_2, \ldots, F_k$ such that, for $1 \leq i \leq k$, any two vectors in $F_i$ are compatible. (two vectors $f_i, f_j$ are compatible if they do not differ at any position or if for any $f_i[k] \neq f_j[k]$, we have $f_i[k] = N$ or $f_j[k] = N$.
- **MEASURE:** Cardinality of the partition, i.e., the number of disjoint subsets, to be minimised.
Example – CMV(ρ)

\[ f_1 = 1NN01 \]
\[ f_2 = N10N1 \]
\[ f_3 = 1N0N1 \]
\[ f_4 = N1N11 \]
\[ f_5 = NN011 \]
\[ f_6 = 0N1N1 \]
\[ f_7 = 00N1N \]
Talk Overview

- Background
- Problem statement
- The NP-hardness of CMV(2)
- Approximation of CMV(p)
- Conclusion
Background

- DNA Microarrays
- Clone
- Probe
- Hybridisation Process
- Fingerprint of a clone
Background – DNA Microarrays

An orderly arrangement of DNA samples on a single chip

Labeled DNA/RNA mixture flushed over array of short DNA fragments

Laser activation of fluorescent labels
Background – Clone & Probe

- **Clone**: A type of DNA sequence
- **Probe**: A short DNA Fragments

For example:

**Clone**  
...TGTGCTTGGCTAGATAGATGC...

**Probe**  
CTTGGCTAGATAGA
Background – Hybridisation Process

Fingerprints

\[
\begin{align*}
0, 1, \ldots & 0 \\
1, N, \ldots & 0 \\
\vdots & \\
0, 1, \ldots & N
\end{align*}
\]
Fingerprint of a clone

A vector consisting of the hybridisation intensity values between the clone and each probe –

1: hybridisation
0: no hybridisation
N: unknown

<table>
<thead>
<tr>
<th>Clone</th>
<th>P_1</th>
<th>P_2</th>
<th>...</th>
<th>P_l</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_1</td>
<td>N</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>C_2</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_i</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C_n</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

C_1 and P_2 did not hybridise
Missing value
C_i and P_l hybridised
Fingerprint vector of C_n
Problem statement
Definitions – Compatible

Given a fingerprint vector set $F = \{f_1, f_1', \ldots, f_n\}$ where $f_i$ is a 0-1-N vector of length $L$, $f_i$ and $f_j$ are compatible if they do not differ at any position or if for any $f_i[k] \neq f_j[k]$, we have $f_i[k] = N$ or $f_j[k] = N$.

For example:

- $f_1 : 101N$
- $f_2 : 1N10$
- $f_3 : N110$

$f_1$ and $f_2$ are compatible

$f_2$ and $f_3$ are compatible

$f_1$ and $f_2$ are NOT compatible
Definitions – Resolved Vector

A 0-1 vector $r$ is a resolved vector of a 0-1-N vector $f$ if $r[i]=f[i]$ for all $1 \leq i \leq \mathcal{U}$ such that $f[i]=1$ or $f[i]=0$.

Let $\text{res}(f)$ be the set of all resolved vectors compatible with $f$.

For example:

$F = \{f_1, f_2\} = \{0N1, N01\}$

$\text{res}(f_1) = \{001, 011\}$

$\text{res}(f_2) = \{001, 101\}$

001 is a resolved vector for both $f_1$ and $f_2$. 
Definition – CMV($p$)

- **INSTANCE:** A set $F$ of 0-1-N fingerprint vectors with at most $p$ missing values per vector.
- **FEASIBLE SOLUTION:** A partition of $F$ into disjoint subsets $F_1, F_2, \ldots, F_k$ such that, for $1 \leq i \leq k$, any two fingerprint vectors in $F_i$ are compatible.
- **MEASURE:** Cardinality of the partition, i.e., the number of disjoint subsets, to be minimised.
Example – CMV($\rho$)
Computation Complexity

- The problem CMV\((p)\) is NP-hard, for any \(p \geq 2\).
- The problem CMV\((1)\) can be solved in polynomial time.
- For any \(p\), CMV\((p)\) can be approximated with ratio \(\min(1 + \ln n, 2 + p\ln l)\) in \(O(nl2^p)\) time.
- CMV\((2)\) is APX-hard.
Definitions – IECMV\((p)\)

Inside compatible clustering with \(p\) missing values (IECMV\((p)\)): The number of compatible pairs of vectors within the same clusters is maximised, i.e., we want to find a partition \(P\) maximising the sum of \(\sum_{i=1}^{k} (\frac{|P_i|}{2})\)

IECMV\((p)\) is APX-hard and can be approximated with ratio 2
Example – IECMV($p$)

To find a partition $P$ maximising the sum of $\sum_{i=1}^{k} (\frac{|P_i|}{2})$

\[
\sum_{i=1}^{k} (\frac{|P_i|}{2}) = 3
\]

\[
\sum_{i=1}^{k} (\frac{|P_i|}{2}) = 5
\]
Definitions – OECMV$(p)$

Outside compatible clustering with $p$ missing values (OECMV$(p)$): The number of compatible pairs of vectors belonging to different clusters is minimised.

OECMV$(p)$ is APX-hard and can be approximated with ratio $1/2$. 
Example – IECMV\((p)\)

To find a partition \(P\) minimising the number of compatible pairs of vectors belonging to different clusters.

\[
\sum = 16
\]

\[
\sum = 12
\]
The NP-hardness of CMV(2)
Minimum Vertex Cover

**INSTANCE:** Given a graph $H = (V, E)$

**SOLUTION:** Find a subsets $V' \subseteq V$ such that, for each edge $(u, v) \in E$, at least one of $u$ and $v$ belongs to $V'$

**MEASURE:** Cardinality of $V'$, to be minimised.

Minimum vertex cover problem on planar, cubic, 3-connected, triangle-free graphs is known to be NP-hard (Uehara 1996)
Planar graph

A graph is planar if it can be drawn in a plane without graph edges crossing (i.e., it has graph crossing number 0).
Cubic Graph

Cubic graphs, also called trivalent graphs, are graphs all of whose nodes have degree 3 (i.e., 3-regular graphs).
A graph is said to be 3-connected if there does not exist a set of 2 vertices whose removal disconnects the graph, i.e., the vertex connectivity of is (Skiena 1990, p. 177).
Triangle-free Graph

A triangle-free graph is a graph containing no graph cycles of length three.
The NP-hardness of CMV (2)

Let $e \in E$ be an arbitrary edge incident with faces $A$ and $B$. Denote the faces incident with the ends of $e$ as $A$, $B$, $X$ and $A$, $B$, $Y$ respectively.
The NP-hardness of CMV (2)

Let $F$ denote the set of faces in $G$. For each edge in $G$, define a set of 0-1-N fingerprint vector $f_e$ of length $|F|$ as follow:

- Set the positions $A$ and $B$ to 1
- Set the positions $X$ and $Y$ to $N$
- Set the remaining ones to 0
The NP-hardness of CMV (2)

Lemma 1 Given a planar, cubic, 3-connected, triangle-free graph. Edges $e$ and $e'$ share a common vertex if and only if the vectors $f_e$ and $f_{e'}$ are compatible.
The NP-hardness of CMV (2)

Proof:
First, suppose that the edges $e$ and $e'$ share a common vertex $v$. Let the faces incident with $v$ be denoted as $P$, $Q$, $R$.

$$
\begin{array}{cccc}
P & Q & R & X \\
\hat{f}_e & 1 & 1 & N & N & 0 \\
\hat{f}_{e'} & 1 & N & 1 & 0 & N
\end{array}
$$
The NP-hardness of CMV (2)

Proof:
Second, suppose that for some non-incident edges $e$ and $e'$, the vectors $f_e$ and $f_{e'}$ are compatible. Therefore, $\{A, B\} \subset \{A', B', X', Y'\}$ and $\{A', B'\} \subset \{A, B, X, Y\}$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A'</th>
<th>B</th>
<th>B'</th>
<th>X</th>
<th>X'</th>
<th>Y</th>
<th>Y'</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_e$</td>
<td>1</td>
<td>1</td>
<td></td>
<td>N</td>
<td></td>
<td>N</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_{e'}$</td>
<td>1</td>
<td>1</td>
<td></td>
<td>N</td>
<td></td>
<td>N</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The NP-hardness of CMV (2)

\{A, B\} \subset \{A', B', X', Y'\} \text{ and } \{A', B'\} \subset \{A, B, X, Y\}

1. A = A', B = X', B' = X
2. A = A', B = B'
3. A = X', B = Y', A' = X, B' = Y
The NP-hardness of CMV (2)

1. $A = A', B = X', B' = X$
The NP-hardness of CMV (2)

2. $A = A'$, $B = B'$
The NP – hardness of CMV (2)

3. \( A = X', B = Y', A' = X, B' = Y \)
The NP-hardness of CMV (2)

It remains to show that the vectors can be divided into $k$ clusters if and only if $G$ has a vertex cover of cardinality $k$. 
The NP-hardness of CMV (2)

Proof:
First, suppose that the vectors have been divided into clusters $F_1, F_2, \ldots F_k$. As every pair of vectors from $F_i$ share a common vertex and there are no triangles in $G$. Therefore, all the vectors from $F_i$ share a common vertex.
The NP-hardness of CMV (2)

Suppose that the vectors have been divided into clusters $F_1, F_2, \ldots F_k$. 
The NP-hardness of CMV (2)

Proof:
Second, suppose that we have a vertex cover $u_1, u_2, \ldots, u_k$ for G. Clearly, we can divide the set of vectors into $k$ clusters selecting to the $i$-th cluster the vectors corresponding to the edges incident with $u_i$. 
Approximation of CMV ($p$)

Theorem 2 CMV($p$) can be approximated in time $O(nl2^p)$ with ratio $\min(1 + \ln n, 2 + pln l)$. For $p = O(\log n)$ the approximation algorithm runs in polynomial time.
Approximation of CMV \((p)\)

Proof: Apply the greedy heuristic which iterates the following step while the set of remaining vectors is non-empty:

- add the largest possible cluster to the current clustering
- remove the vectors belonging to the cluster from the set of remaining vectors.

Since the greedy heuristic can be interpreted as covering the input set of vectors with subsets in one-to-one correspondence with the possible maximal clusters, it yields the \(\min(1 + \ln n, 2 + p \ln l)\) approximation ratio (Johnson, ‘Approximation algorithms for combinatorial problems’, 1974)
Running Time of CMV \((p)\)

For a fingerprint \(x\) (set of fingerprints \(X\)), let \(\text{res}(x)(\text{res}(X))\) be the set of all resolved fingerprints compatible with \(x\) (with some elements of \(X\)).
Running Time of CMV \((p)\)

- \(x = \) a fingerprint vectors
- \(\text{res}(x) = \) resolved vectors of fingerprint vector \(x\)
- \(X = \) a set of fingerprint vectors
- \(\text{res}(X) = \) resolved vectors of some elements of \(X\)

\[
\begin{align*}
x_1 &= 0N1 \\
x_2 &= N01 \\
\text{res}(x_1) &= \{001, 011\} \\
\text{res}(x_2) &= \{001, 101\} \\
X &= \{x_1, x_2\} = \{0N1, N01\} \\
\text{res}(X) &= \{001, 011, 101\}
\end{align*}
\]
Running Time of CMV ($p$)

First, we compute an auxiliary bipartite graph $H$ with set of vertices $(A, B)$ where $A = res(F)$ and $B = F$. For $x \in A$ and $y \in B$, the edge $xy$ is present in $H$ iff $x$ and $y$ are compatible.
Running Time of CMV ($p$)

The algorithm is summarised below:

```
Algorithm Construction of $H = (A, B, E)$
1  $A := \emptyset$
2  $B := F$
3  $E := \emptyset$
4  for all $x \in B$ do
4.1  for all $y \in res(x)$ do
4.1.1  if $y \notin A$ then
4.1.1.1  Insert($y, A$)
     endif
4.1.2  Insert($E, xy$)
     endfor
5  endfor
End Construction of $H = (A, B, E)$
```

The construction takes $O(nl2^p)$ time
Running Time of CMV ($p$)

Algorithm  Greedy Clustering

1. for $i := 1$ to $n$ do
   1.1 $Q_i := \emptyset$
   endfor
2. for all $x \in A$ do
   2.1 Insert($x, Q_{\text{deg}(x)}$)
   endfor
3. for $i := n$ to 1 do
   3.1 while $Q_i$ is not empty do
      3.1.1 $x := \text{Delete}(Q_i)$
      3.1.2 Begin reporting a new cluster
      3.1.3 for all $y$ neighbor of $x$ do
         3.1.3.1 Report($y$)
         3.1.3.2 Delete($y$)
      endfor
      3.1.4 Delete($x$)
   endwhile
endfor
End  Greedy Clustering

- $Q_i$: Stores all the vertices from $A$ with degree exactly $i$.
- $\text{Delete}(x)$: Removes from $H$ vertex $x$, all the edges incident with $x$ and move the neighbours of $x$ to the relevant queues.
- As every edge and every vertex is deleted from $H$ only once, it is clear the complexity of the phase is linear in the size of $H$, i.e.,
  \[ O(|H|) = O(|A| |B|) = O(n2^p) \]
Conclusion

- The problem CMV($p$) is NP-hard, for any $p \geq 2$.
- CMV(2) is APX-hard.
- CMV($p$) can be approximated with ratio $\min(1 + \ln n, 2 + pl\ln l)$ in $O(nl2^p)$ time.
- IECMV($p$) is APX-hard and can be approximated with ratio 2.
- OECMV($p$) is APX-hard and can be approximated with ratio $1/2$. 


THANK YOU!

QUESTIONS?
Minimum Vertex Cover

Let $S$ be a collection of subsets of a finite set $X$. The smallest subset $Y$ of $X$ that meets every member of $S$ is called the vertex cover. Some authors call any such set a vertex cover, and then refer to the minimum vertex cover (Skiena 1990, p. 218). Finding the minimum vertex cover is an NP-complete problem.