On Mining Disjunctive Closed Itemsets in Microarray Gene Expression Data

Supplementary Material – Proofs for propositions of Section "Basic concepts"

Renato Vimieiro and Pablo Moscato

Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine

The University of Newcastle

Hunter Medical Research Institute

Lot 1, Kookaburra Circuit, New Lambton Heights, NSW, 2305, Australia.

Email: {renato.vimieiro, pablo.moscato}@newcastle.edu.au

Proposition 1. *Let I, I_1, I_2 be arbitrary itemsets, and let O, O_1, O_2 be arbitrary sets of samples in a data set (S, F, R). Functions α, β and γ have the following properties:*

1. $I \subseteq \gamma(I)$;
2. $\alpha(\beta(O)) \subseteq O$;
3. *if* $I_1 \subseteq I_2$, *then* $\alpha(I_1) \subseteq \alpha(I_2)$;
4. *if* $O_2 \subseteq O_1$, *then* $\beta(O_2) \subseteq \beta(O_1)$;
5. $\alpha(I) = \alpha(\gamma(I))$;
6. $\beta(O) = \gamma(\beta(O))$;

Proof. While Vimieiro and Moscato (2012) have already demonstrated items 1, 3 and 5, they have not demonstrated the remaining properties. We provide the formal proof for those properties here.

(2) Let $s \in \alpha(\beta(O))$ be an arbitrary sample. Since $s \in \alpha(\beta(O))$, we know by the definition of α that there exists $i \in \beta(O)$ such that $(s, i) \in R$. By the definition of β, we know that $i \in \beta(O)$ if and only if $\alpha(i) \subseteq O$. Given that $(s, i) \in R$, we know that $s \in \alpha(i)$. Therefore, $s \in O$. Since s is arbitrary, it follows that $\alpha(\beta(O)) \subseteq O$.

(4) Let us suppose that $O_2 \subseteq O_1$. Now, let $i \in \beta(O)$ be an arbitrary feature. By the definition, we have that $\alpha(i) \subseteq O_2$. Thus, $\alpha(i) \subseteq O_1$, because we supposed that $O_2 \subseteq O_1$. Since i is arbitrary, it follows that $\beta(O_2) \subseteq \beta(O_1)$. Therefore, if $O_2 \subseteq O_1$, then $\beta(O_2) \subseteq \beta(O_1)$ as required.

(6) $\beta(O) \subseteq \gamma(\beta(O))$ follows straight from item 1, and $\beta(O) \supseteq \gamma(\beta(O))$ follows from item 4.

□

Proposition 2. *Let $D = (S, F, R)$ be an arbitrary data set. Let $A = \{\beta(O) \mid O \subseteq S\}$ be the family of all sets of features describing each set of samples, and let $DCI = \{\gamma(I) \mid I \subseteq F\}$. Then, $A = DCI$.*

Proof.

($\subseteq$) Let X be an arbitrary subset of samples. By definition, $\beta(X) \in A$. From Proposition 1 property 6 , we have $\beta(X) = \gamma(\beta(X))$. Therefore $\beta(X) \in DCI$.

($\supseteq$) Let Y be an arbitrary closed itemset from DCI. We know from the definition of α that $\alpha(Y) \subseteq S$. Thus $\beta(\alpha(Y)) \in A$ by definition. But, since Y is closed, we know that $Y = \beta(\alpha(Y))$. Therefore $Y \in A$.

$\square$

Proposition 3. *Let X and Y be two arbitrary sets of samples such that $Y \subseteq X$. Then, $TT|_Y \subseteq TT|_X$.*

Proof. Follows straight from the definitions of transposed conditional tables and function β, and property 4 of Proposition 1. $\square$

Proposition 4. $\beta(X) = \{f \mid (f, \alpha(f)) \in TT|_X\}$

Proof. This proposition follows straight from the definitions of β and $TT|_X$. $\square$

Proposition 5. *Let $X \subseteq S$ be an arbitrary subset of samples of a data set (S, F, R). Then, X is closed if and only if $X = \bigcup\{\alpha(f) \mid (f, \alpha(f)) \in TT|_X\}$. In other words, $\alpha(\beta(X)) = \bigcup\{\alpha(f) \mid (f, \alpha(f)) \in TT|_X\}$.*

Proof. Proposition 4 allows us to restate $\bigcup\{\alpha(f) \mid (f, \alpha(f)) \in TT|_X\}$ as $\bigcup\{\alpha(f) \mid f \in \beta(X)\}$. Then, our target becomes proving $\gamma(X) = \bigcup\{\alpha(f) \mid f \in \beta(X)\}$. Thus, We must consider two cases:

$\subseteq$ Let $s \in \alpha(\beta(X)) = \gamma(X)$ be an arbitrary sample. By the definition of α, we know that there exists $i \in \beta(X)$ such that $(s, i) \in R$. Thus, $s \in \alpha(i)$ and, therefore, $s \in \bigcup\{\alpha(f) \mid f \in \beta(X)\}$. Since s is arbitrary, it follows that $\gamma(X) \subseteq \bigcup\{\alpha(f) \mid f \in \beta(X)\}$.

$\supseteq$ Conversely, let $s \in \bigcup\{\alpha(f) \mid f \in \beta(X)\}$ be an arbitrary sample. Then, there exists $f \in \beta(X)$ such that $s \in \alpha(f)$. By the definition of α, we know that $(s, f) \in R$. Since $f \in \beta(X)$, it follows that $s \in \alpha(\beta(X))$. Therefore, $\bigcup\{\alpha(f) \mid f \in \beta(X)\} \subseteq \alpha(\beta(X)) = \gamma(X)$, because s is an arbitrary sample.

$\square$

Proposition 6 (Prune 1). *Let X and Y be two arbitrary sets of samples such that Y is a son of X in the enumeration tree — Y is a subset of X and $|Y| = |X| - 1$. Let $i \in X$ be such that $Y = X - \{i\}$, i.e., i is the element removed from X to obtain Y in the enumeration process (more details in Section 3 of the main text).* The branch rooted by Y can be safely pruned *whenever there exists $s \succ i$ (s greater than i considering an arbitrary total order) such that $s \in Y$ and $s \notin \bigcup\{\alpha(f) \mid (f, \alpha(f)) \in TT|_Y\}$.*

Proof. The enumeration process is carried out in order. Thus, only elements that are less than i can be removed from Y. Since no feature in $TT|_Y$ contains the sample s, clearly, Y is not closed because of Proposition 5. Moreover, conforming to Proposition 3, none of the conditional tables derived from subsets of Y include s in any of their features. The sample s, by the other hand, will be present in all subsets of Y, because of the enumeration process itself. Therefore, none of the subsets of Y obtained through the enumeration process is a closed set and, then, the whole branch can be trimmed off. $\square$

Proposition 7 (**Reducible sets**)**.** *Let X be the current set of samples being processed in the enumeration tree. Let i be the last element removed from the candidate set X in the enumeration process (see Section 3 of the main text for details). We say that X is* reducible *if X contains elements lower than i that do not occur in any feature in the conditional table of X. More formally, X is reducible if the set $Reducible = \{j \in X \mid j \prec i \wedge j \notin \bigcup\{\alpha(f) \mid (f, \alpha(f)) \in TT|_X\}\}$ is not empty. If X is reducible — there are elements in X that do not occur in any feature of $TT|_X$—, then we can exclude all of them from X at once, and reset $X - Reducible$ as the root of the current subtree in the enumeration process. After that, we carry on with the process normally from the new root $X - Reducible$.*

Proof. Let $j \in Reducible$. We face two situations in the enumeration:

1. We keep j in X and remove the next element k from X. We know that j does not belong to any feature in $TT|_{X-\{k\}}$ and thus this new branch is pruned by Proposition 6. Note that this situations happens for any $j \in Reducible$.

2. We remove j from X. In this case, the conditional table of $Y = X - \{j\}$ is identical to $TT|_X$ and this new node still does not represent a closed set.

Both situations described above corroborates the uselessness of the items in *Reducible* for the enumeration process in the current subtree. Therefore, all of them may be removed at once from X, and the process restarted from this new node. □

Proposition 8 (**Prune 2**)**.** *Let X be an arbitrary set of samples, and $TT|_X$ its transposed conditional table. If $|X| \leq min\{|\alpha(f)| \mid (f, \alpha(f)) \in TT|_X\}$, then $TT|_Y = \emptyset$ for every $Y \subset X$.*

Proof. We see that if the size of X is at most the size of the smallest set of samples associated with a feature in the conditional table, then there is no feature able to describe any subset of X, i.e. there is no feature such that the set of samples is a subset of a subset of X. □

References

Vimieiro, R. and P. Moscato (2012). Mining disjunctive minimal generators with TitanicOR. *Expert Systems with Applications 39*(9), 8228–8238.