International Journal of Pure and Applied Mathematics
Volume 78 No. 2 2012, 245-251
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu

ijpam.eu

MAXIMAL FREQUENT SETS

Eungchun Cho
Kentucky State University
Frankfort, KY, 40601, USA

Abstract: Given a collection $\mathcal{U}$ of subsets of a finite set X, subsets of X that are covered by many sets in $\mathcal{U}$ are called frequent sets. Frequent sets are of interest in many applications, statistics, machine learning, pattern recognition and data mining. Characterizations of maximal frequent sets in terms of equivalence relation and partial ordering, and an algorithm of finding them are given.

AMS Subject Classification: 06A06, 62D05, 03E75
Key Words: frequent set, partial order, pattern recognition, data mining, machine learning

1. Introduction

The concept of frequent sets introduced in 1993 by Agrawal et al. [1] [2] [3] has received a great interest among researchers in pattern recognition, data mining and machine learning, for example, [4]. Extensive references and survey papers from researchers in application to pattern recognition and data mining are readily available, we refer [4] and only give a brief list of reference of the sminal papers of the idea. In this paper, we present a simple description of maximal frequent sets employing basic set theoretic concepts such as equivalence relation and partial ordering.

Let X be a finite set and $\mathcal{U}$ a set of subsets of X. A k-frequent set is a subset of X that is contained in (exactly, not more than) k subsets in $\mathcal{U}$. We are interested in k-frequent sets for large k. Checking the frequency of all subsets of X to find k-frequent sets is infeasible when X and $\mathcal{U}$ are not small.

Received: April 7, 2012

If, however, many elements of X are covered together by groups of sets in $\mathcal{U}$, the complexity is reduced by considering the quotient space of X with respect to the membership relation. We characterize maximal k-frequent sets, which suggests an efficient algorithm for finding all maximal k-frequent sets.

2. Notation

Following notation and abbreviations are used.

$$\begin{aligned}
\mathcal{P}(X)[\mathcal{P}(\mathcal{U})resp.] &: \text{ the power set of } X \text{ [}\mathcal{U} \text{ resp.]} \\
\bigcup B[\bigcap Bresp.] &= \bigcup_{S\in B} S[\bigcap_{S\in B} Sresp.] \\
|A| &: \text{ the number of elements in } A. \\
[x] &= \{y \in X : y \sim x\} \\
\hat{X} &= X/\sim \\
f(x) &= \{S \in \mathcal{U} : x \in S\} \\
\hat{f}([x]) &= \{S \in \mathcal{U} : [x] \subset S\} \\
g(A) &= \{S \in \mathcal{U} : A \subset S\} \\
h(A) &= \bigcap g(A) \\
fr(A) &= |g(A)| : \text{ the frequency of } A. \\
F_k &= \{A \subset X : fr(A) = k\} \\
MF_k &: \text{ the set of all maximal k -frequent sets.} \\
C_k &= \{[x] : fr([x]) = k\}
\end{aligned}$$

3. Definitions and Lemmas

In this section, a map from X into $\mathcal{P}(\mathcal{U})$ and its extensions, an equivalence relation on X, a partail order on the quotient space $\hat{X}$, the frequency and maximal frequent sets will be defined.

Let $X = \{x_1, \ldots, x_m\}$ and $\mathcal{U} = \{S_1, \ldots, S_n\} \subset \mathcal{P}(X)$ such that $\bigcup \mathcal{U} = X$. Define a map $f : X \to \mathcal{P}(\mathcal{U})$ by

$$f(x) = \{S \in \mathcal{U} : x \in S\} \tag{1}$$

The cardinality of the set $f(x)$ will be called the frequency of x. f defines an equivalence relation $\sim$ on X:

$$x \sim y \quad \text{iff} \quad f(x) = f(y) \tag{2}$$

p denotes the projection map of X onto the quotient space

$$X/\sim = \{[x] : x \in X\} \tag{3}$$

In the following, $X/\sim$ will be denoted by $\hat{X}$. There is a unique injective map $\hat{f}$ from $\hat{X}$ into $\mathcal{P}(\mathcal{U})$ such that $f = \hat{f} \circ p$, namely, the map defined by

$$\hat{f}([x]) = \{S : [x] \subset S\} \tag{4}$$

Define a map $g : \mathcal{P}(X) \to \mathcal{P}(\mathcal{U})$ by

$$g(A) = \{S \in \mathcal{U} : A \subset S\}, \tag{5}$$

which generalizes both f and $\hat{f}$:

$$g(\{x\}) = \hat{f}([x]) = f(x) \tag{6}$$

The frequency of a subset A is defined as the cardinality of $g(A)$, the number of sets in $\mathcal{U}$ that contain A:

$$fr(A) = |g(A)| \tag{7}$$

Since

$$fr(x) \quad = \quad fr([x]), \quad fr(A) = fr(\bigcup p(A))$$

the frequency can be defined at the equivalence class level. C_k denotes the set of all k-frequent equivalence classes:

$$C_k = \{[u] : fr([u]) = k\} \tag{8}$$

Following propositions are easily verified.

Proposition 1. *If $A \subset B$, then $g(B) \subset g(A)$.*

Proposition 2. $g(\bigcup_i A_i) = \bigcap_i g(A_i))$.

Proposition 3. $g(A) = \bigcap \{f(x), x \in A\} = \bigcap \{\hat{f}([x]) : x \in A\}$.

A partial order $\leq$ on X is induced by the partial order $\subset$ on $\mathcal{P}(\mathcal{U})$ via f:

$$x \leq y \quad \text{iff} \quad f(x) \subset f(y) \tag{9}$$

$\hat{X}$ inherits the partial oreder:

$$[x] \leq [y] \quad \text{iff} \quad \hat{f}([x]) \subset \hat{f}([y]) \tag{10}$$

Define a map $h : \mathcal{P}(X) \to \mathcal{P}(X)$ by

$$h(A) = \bigcap \{S \in \mathcal{U} : A \subset S\} \tag{11}$$

Let F_k denote the set of all k-frequent sets. $A \in F_k$ is maximal means $A \subset B \in F_k$ implies $A = B$. MF_k denotes the set of all maximal k-frequent sets.

We have the following

Lemma 1. $h^2 = h$

Proof. Since $g(h(A)) = g(A)$, $h^2(A) \subset h(A)$. *Obviously,* $A \subset h(A)$ *and* $h(A) \subset h^2(A)$.

Lemma 2. $h(\{u\}) = h([u]) = h(\bigcup \{[v] : [u] \leq [v]\} = \bigcup \{[v] : [u] \leq [v]\}$.

Proof. If $\hat{f}([u]) \subset \hat{f}([v])$ *then* $[v] \subset h([v]) \subset h([u])$. *This implies* $\bigcup \{[v] : \hat{f}([u]) \subset \hat{f}([v])\} \subset h([u])$.

Corollary 1. *If* $[u]$ *is maximal with respect to* $\leq$, *then* $h([u]) = [u]$.

4. Main Theorems, Algorithm and Example

In this section, main results, characterization of the set MF_k of all maximal k-frequent sets, algorithm for building MF_k, and a simple example are given. We state main theorems first, which are proved by the following lemmas. The main result below shows any maximal frequent sets is a union of equivalence classes, that is, maximal frequent sets can be constructed on $\hat{X}$ instead of X.

Theorem 1. *Any maximal* k*-frequent set is a union of equivalence classes* $[x]$.

Theorem 2. *If $[u] \in C_k$ is maximal with respect to $\leq$, then $h([u]) = [u]$ and $[u] \in MF_k$.*

In other words, any maximal (with respect to $\leq$ on $\hat{X}$) k-frequent equivalence class is a maximal k-frequent set. Every subset of a k-frequent set is obviously m-frequent for $m \geq k$. However, maximal k-frequent equivalence classes are pure maximal k-frequent sets in the sense that they are maximally frequent and $fr(A) = k$ for every $A \subset [u]$.

Theorem 3. *$\bigcup_i [u_i]$ is $|\bigcap_i g[u_i])|$-frequent.*

Proof. Let $A \in C_k$ be maximal. Then $g(A) = k$ and $g(A) \subset g([x])$ for every $x \in A$. Then $g(A) = \bigcap_{x \in A} g([x])$. Since $E \bigcup \{\emptyset\}$ is closed under intersection, $g(A) = g([u])$ for some $[u] \in C_k$.

Some lemmas are stated without proof as they are obvious consequences of the definitions.

Lemma 3. *If $[x] \in C_k$ then $h([x]) \in MF_k$.*

Proof. Since $g(h([x])) = \hat{f}([x])$, $h([x])$ is k-frequent. Any k frequent superset B of $h([x])$ must have $g(B) = g(h([x]))$, hence $B = h([x])$.

Lemma 4. *For any nonempty $A \subset X$, $h(A) = \bigcup_{x \in A}[x]$ and $h(A) \in MF_k$ where $k = min(fr(x) : x \in A)$.*

Lemma 5. *If $A \subset X$ and $B \subset A$ is the set of all minimal elements of A, i.e., $x \in B$, $y \in A$ and $y \leq x$ then $x = y$, then $h(A) = h(B)$.*

Proposition 4. *Let $A \in MF_k$, then there exists at most one $[x] \subset A$ that belongs to C_k.*

Proof. If $[x], [y] \subset A$ and both belong to C_k, then $g(A) \subset \hat{f}([x]) \bigcap \hat{f}([y])$ and $|g(A)| < k$.

Proposition 5. *$A, B \in MF_k$, and $A \bigcap B \neq \emptyset$. Then $A = B$.*

Proposition 6. *Let $A \in MF_k$, and there exist at least one $x \in A$ with $fr(x) = k$ then $A = h([x])$.*

Proof. Since $[x] \in C_k$, $h([x]) \in MF_k$.

Proposition 7. *If $A, B \in F_k$ and $A \subset B$, the $g(A) = g(B)$.*
Proof. $A \subset B$ implies $g(B) \subset g(A)$. Since $|g(A)| = |g(B)|$, $g(A) = g(B)$.

Now we state an algorithm for finding maxiaml frequent sets. It was shown $\{\bigcup_{[x]\leq[u]}[u] : [x] \in C_k\} \subset MF_k$ and any $A \in MF_k$ contains at most one $[x] \in C_k$. If there exist such $[x]$ then $A = h([x])$. Therefore, MF_k is found by the following procedures:

1. Construct the incidence matrix M representing $\in$: $M_{ij} = x_i \in S_j$.
2. Find $f(x)$ for each $x \in X$.
3. Cnstruct $\hat{X}$.
4. Cnstruct C_k.
5. Construct the incidence matrix representing $\leq$.
6. Take $h([u])$ for each $[u] \in C_k$, which is a maximal k-frequent set containing $[u]$.

If every $A \in MF_k$ contains one $[x] \in C_k$, then $MF_k = h(C_k)$. In generl, $h(C_k)$ is a subset of MF_k. In other words, there may exist maximal k-frequent sets that are not found by this algorithm unless the condition is satisfied. In practical applications, large subsets with high frequency are of interest. Therefore, we consider the set M of $[x]$ that are maximal with respect to $\leq$ on $\hat{X}$ and take $[u] = \text{argmax}(fr([x], [x] \in M)$ and $k = fr([u])$. Then $h([u]) = [u]$ and it is a maximal k-frequent set with the higest possible frequency k.

Example 1.

$$\begin{array}{lcl} X & = & \{1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ \} \\ S_1 & = & \{1\ \ \ \ 4\ 5\ \ \ 8\ \} \\ S_2 & = & \{1\ 2\ \ \ 4\ \ 6\ \ \ 8\ \} \\ S_3 & = & \{\ \ \ 3\ 4\ 5\ \ 7\ 8\ \} \end{array}$$

and $\mathcal{U} = \{S_1, S_2, S_3\}$.

$$\begin{array}{lcl} \hat{X} & = & \{[1], [2], [3], [4], [5]\} \\ C_1 & = & \{[2], [3]\} \\ C_2 & = & \{[1], [5]\} \end{array}$$

$$
\begin{aligned}
C_3 &= \{[4]\} \\
MF_1 &= \{S_1, S_2, S_3\} \\
MF_2 &= \{\{1,4,8\},\{4,5,8\}\} \\
MF_3 &= \{\{4,8\}\}
\end{aligned}
$$

Acknowledgments

Eungchun Cho's work at Seoul National University was supported by The Korea Research Foundation and The Korean Federation of Science and Technology Societies Grant funded by Korea Government (MOEHRD, Basic Research Promotion Fund).

References

[1] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, In: *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (Ed-s: P. Buneman, S. Jajodia); *SIGMOD Record*, ACM Press, **22**, No. 2 (1993), 207-216.

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A. Verkamo, Fast discovery of association rules, In: *Advances in Knowledge Discovery and Data Mining* (Ed-s: U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy), MIT Press (1996), 307-328.

[3] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, In: *Proceedings 20-th International Conference on Very Large Data Bases* (Ed-s: J. Bocca, M. Jarke, C. Zaniolo), Morgan Kaufmann (1994), 487-499.

[4] S.E Brossette, A.P. Sprague, Frequent Sets in the presence of clones: An example from medical serveillence, *Discrete Mathematical Problems with Medical Applications Series in Discrete Mathematics and Theoretical Computer Science*, **55** (1999), 165-170.