

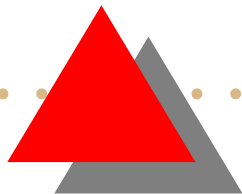


*An Inclusion-Exclusion Result for  
Boolean Polynomials and Its  
Applications in Data Mining*

Szymon Jaroszewicz<sup>\*</sup>, Dan A. Simovici<sup>\*</sup>, and Ivo Rosenberg<sup>\*\*</sup>

<sup>\*</sup>University of Massachusetts at Boston

<sup>\*\*</sup> Université de Montréal



# Market Basket Data

customer ID	beer	bread	...	diapers
101	1	0	...	1
103	0	1	...	1
107	1	1	...	1
...	...	...	...	...

- **Items:** binary attributes
- **Itemsets:** sets of items

# Frequent Itemsets

- **Support** of an itemset  $I$  in relation  $\rho$ :

$$\text{supp}_{\rho}(I) = \frac{|\{t \in \rho : t[I] = (1, \dots, 1)\}|}{|\rho|}$$

(essentially the same as probability)

- Itemset  $I$  is frequent if

$$\text{supp}(I) > \text{minsupp}$$

- `Apriori` algorithm efficiently finds all frequent itemsets



# *Inspiration*

H. Manilla et al. [1996,2001]: Use frequent itemsets to get support of arbitrary queries, e.g.:

$$\text{supp}(\bar{A}\bar{B}) = 1 - \text{supp}(A) - \text{supp}(B) + \text{supp}(AB)$$

(inclusion-exclusion principle)

## *Questions:*

- How to obtain such a formula for arbitrary function?
- Guarantee of accuracy if some supports are unknown?

# *A more general statement*

- Boolean Algebra:  $(B, \bar{\phantom{x}}, \vee, \wedge)$
- Set of variables:  $A = \{a_1, \dots, a_n\}$
- $(A)$  the free Boolean algebra on  $A$  consists of polynomials:
  - $\bar{\phantom{x}}$ ,  $\vee$ , and  $\wedge$  and each  $a_i$  belong to  $(A)$ ;
  - if  $p, q \in (A)$ , then  $\bar{p}$ ,  $(p \vee q)$ ,  $(p \wedge q) \in (A)$ .



# Measures on Boolean Algebras

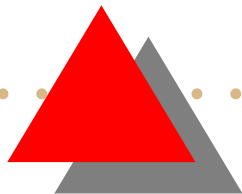
A **measure** on a Boolean Algebra  $(B, , , ^-, \vee, \wedge)$ :  
 $\mu : B \rightarrow [0, \infty]$  s.t.

$$\mu(x \vee y) = \mu(x) + \mu(y)$$

if  $x \wedge y = 0$ .

*Example:*

Support  $\text{supp}$  is a measure on  $(A)$



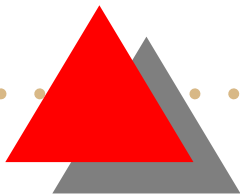



# Representation result

## *Theorem:*

A function  $\mu : (A)$  is a measure  
if and only if

there exists a binary relation  $\rho$ , such that  $\mu(p) =$   
 $\text{supp}_\rho(p)$  for all  $p \in (A)$ .





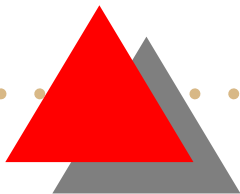
# Question 1 rephrased

For any  $p \in (A)$  and some measure  $\mu$  express  $\mu(p)$  in terms of measures of positive conjunctions

*Examples:*

$$\mu(a_1 \oplus a_2) = \mu(a_1) + \mu(a_2) - 2\mu(a_1 \wedge a_2)$$

$$\mu(\bar{a}_1 \wedge \bar{a}_2) = \mu() - \mu(a_1) - \mu(a_2) + \mu(a_1 \wedge a_2)$$





# *Inclusion-exclusion type result for Exclusive-or*

- $p_1, p_2, \dots, p_m$  are Boolean polynomials
- Let

$$S_k = \sum_{i_1 \leq \dots \leq i_k} \mu(p_{i_1} \wedge p_{i_2} \wedge \dots \wedge p_{i_k})$$

- Then,


$$\mu(p_1 \oplus \dots \oplus p_m) = \sum_{k=1}^m (-2)^{k-1} S_k$$

# Example

*Parity function:*  $a_1 \oplus a_2 \oplus a_3$

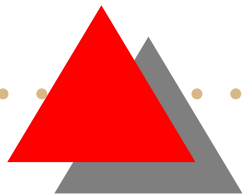
- $S_1 = \mu(a_1) + \mu(a_2) + \mu(a_3)$
- $S_2 = \mu(a_1 \wedge a_2) + \mu(a_2 \wedge a_3) + \mu(a_1 \wedge a_3)$
- $S_3 = \mu(a_1 \wedge a_2 \wedge a_3)$
- giving

$$\mu(a_1 \oplus a_2 \oplus a_3) = S_1 - 2S_2 + 4S_3$$



*Every Boolean polynomial can be represented as exclusive-or of positive conjunctions*

*We can express a measure of any boolean polynomial in terms of measures of positive conjunctions of its variables*



# Bounds

Dropping terms from inclusion-exclusion we get bounds on the measure: **Bonferroni Inequalities:**

$$\sum_{k=1}^{2r} (-2)^{k-1} S_k^\mu \leq \mu(p_1 \oplus \dots \oplus p_m) \leq \sum_{k=1}^{2s+1} (-2)^{k-1} S_k^\mu,$$

for any  $r, s \in$

# Example

- Upper bound:

$$\mu(a_1 \oplus a_2 \oplus a_3) \leq \mu(a_1) + \mu(a_2) + \mu(a_3)$$

- Lower bound:

$$\begin{aligned} \mu(a_1 \oplus a_2 \oplus a_3) &\geq \mu(a_1) + \mu(a_2) + \mu(a_3) \\ &\quad - 2\mu(a_1 \wedge a_2) - 2\mu(a_2 \wedge a_3) - 2\mu(a_1 \wedge a_3) \end{aligned}$$

We can thus obtain bounds for support of any database query

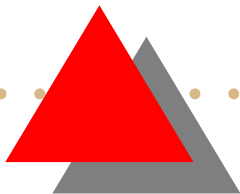


# *There are queries which cannot be approximated*

Parity polynomial:  $p_{par} = a_1 \oplus a_2 \oplus \dots \oplus a_n$

Two relations over  $A = (a_1, a_2, \dots, a_n)$ :

$$\begin{aligned}\rho_{odd} &= \{t \in (A) : n_1(t) \text{ is odd}\}, \\ \rho_{even} &= \{t \in (A) : n_1(t) \text{ is even}\},\end{aligned}$$





# *There are queries which cannot be approximated*

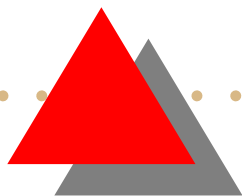
We have:

$$\text{supp}_{\rho_{\text{odd}}}(K) = \text{supp}_{\rho_{\text{even}}}(K) \text{ for all } K \subset A,$$

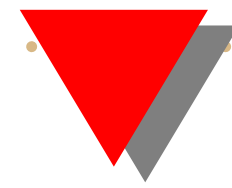
but

$$\text{supp}_{\rho_{\text{odd}}}(p_{\text{par}}) = 1, \text{supp}_{\rho_{\text{even}}}(p_{\text{par}}) = 0$$

One unknown itemset  $A$  can result in huge inaccuracy of  $\text{supp}(p_{\text{par}})$



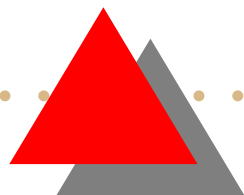
# Tables with missing values



Allow missing values  $(a_i) = \{, , \}$

Define  $\mu$  generalizing support to such tables

- With each attribute  $a_i$  associate a value  $\alpha_i \in [0, 1]$
- If only one attribute  $i$  is missing, multiply tuple's support by  $\alpha_i$
- If more attributes are missing, use independence assumption





# Example

$$\begin{aligned}\mu(\bar{a}_1 \wedge a_2) &= \text{supp}(a_1 = \wedge a_2 =) \\ &+ (1 - \alpha_1) \text{supp}(a_1 = \wedge a_2 =) \\ &+ \alpha_2 \text{supp}(a_1 = \wedge a_2 =) \\ &+ (1 - \alpha_1) \alpha_2 \text{supp}(a_1 = a_2 =)\end{aligned}$$

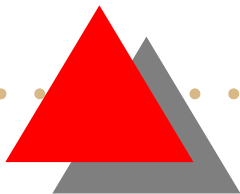


# *Properties of $\mu$*

*Theorem:*  
 $\mu$  is a measure.

*Consequences:*

- $\mu$  gives probabilistically consistent results.
- All previous results apply to  $\mu$



# Example

$a_1$	$a_2$

$$\alpha_2 = 1$$

# Example

[Ragel, Crémilleux 98]: count each itemset where it is defined

$$\text{supp}(a_1) = 0.5 < \text{supp}(a_1 \wedge a_2) = 1$$

[Nayak, Cook 01]: weighted sum of attributes in a row

$$\text{supp}(a_1) = 0.5 < \text{supp}(a_1 \wedge a_2) = 0.75$$

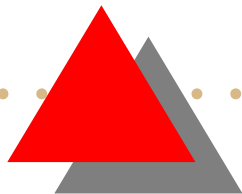
but

$$\mu(a_1) = 0.5 \quad \mu(a_2) = 1 \quad \mu(a_1 \wedge a_2) = 0.5$$



# *Further research*

- More applications in datamining
- Tighter bounds for specific queries
- Case when complete  $S_k$  are not known  
[submitted PKDD'02]





# *Estimates with incomplete $S_k$ 's*

*Main idea:* Apply Bonferroni inequalities recursively

*Example:*

- Known supports:  $A, B, C, AC, BC$
  - Want:  $\text{supp}(ABC)$
1. Estimate  $\text{supp}(AB)$
  2. Use bounds for  $\text{supp}(AB)$  to compute  $S_2$
  3. Compute  $\text{supp}(ABC)$

# *Bonferroni-type inequalities for supports*

The following inequalities hold for any  $t \in$ :

$$\text{supp}(a_1 a_2 \dots a_m) \leq \sum_{k=0}^{2t} (-1)^k m - k - 12t - k S_k$$

$$\text{supp}(a_1 a_2 \dots a_m) \geq \sum_{k=0}^{2t+1} (-1)^{k+1} m - k - 12t + 1 - k S_k$$