

DETECTION OF DISJUNCTIVE ANOMALOUS PATTERNS IN MULTIDIMENSIONAL DATA

Robin Sabhnani

Artur Dubrawski

Jeff Schneider

ML

**Auton
Lab**

Carnegie Mellon

International Society for Disease Surveillance
Eighth Annual Conference, Miami, FL.

Dec 3, 2009

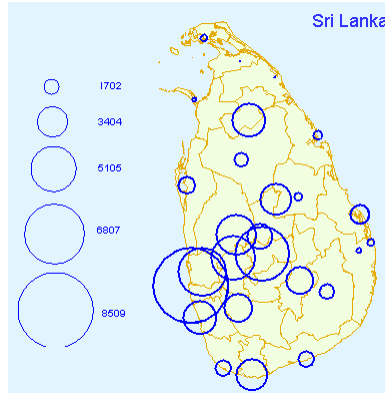
TALK OUTLINE

- Motivation
- Applications
- Multidimensional Data
- Disjunctive Pattern
- Pattern Scoring
- WSARE
- DAD
- Performance evaluation
- Conclusion

MOTIVATION: SRI LANKA DATA

- Sri Lanka Department of Health collects weekly aggregated data across 26 regions and 9 diseases nationwide starting Dec 2006

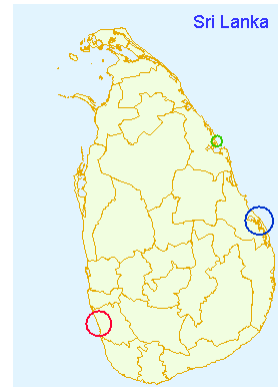
Dengue fever
 Human Rabies
 Dysentery
 Leptospirosis
 Encephalitis
 Typhus fever
 Enteric fever
 Viral Hepatitis
 Food Poisoning



ANALYZE DISEASE: DYSENTERY



Kalutara



Trincomalee

Batticaloa

Kalutata, Trincomalee, Batticaloa

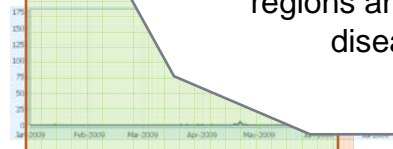


ANALYZE REGION: VAVUNIYA



Dysentery

A disease outbreak can simultaneously affect multiple regions and multiple diseases



Dysentery, Viral hepatitis, Enteric fever

APPLICATION: BIO-SURVEILLANCE

o Data Dimensions

- o Patient visit date
- o Age group
- o Gender
- o Home zip code
- o Chief-complaint

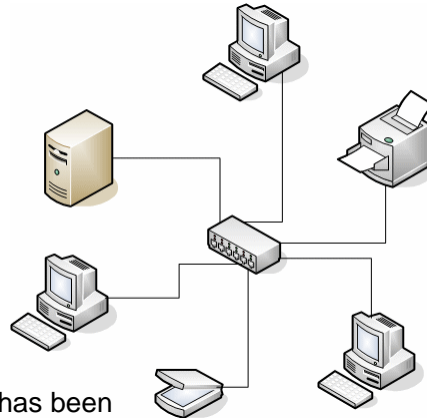


o Example Patterns

- o An influenza disease outbreak among senior citizens in and around the city of Miami, FL.
- o Number of Male patients with respiratory complaints has doubled in the last week as compared to last month

APPLICATION: NETWORK DATA ANALYSIS

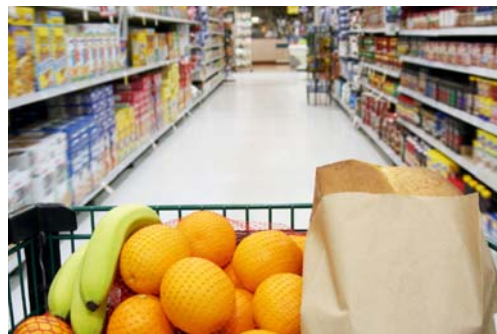
- Data Dimensions
 - Start time
 - Type of activity
 - CPU usage
 - Server load
 - User gender
 - User age group
 - Origination Country



- Example Pattern
 - In the last one hour, there has been a spike in the number of failed logins (Password guessing attack)

APPLICATION: GROCERY CART ANALYSIS

- Data Dimensions
 - Purchase Date
 - Has Fruits?
 - Has Milk?
 - ...
 - Customer gender
 - Customer age group
 - Store location



- Example Patterns
 - Recently, the sale of beer and frozen pizza has quadrupled as compared to the past (Game season)
 - Recently the sale of sweet products out-numbered all other categories (Holiday season)

COMMON DATA CHARACTERISTICS

- Transactional data
- Exactly one temporal dimension (e.g. Date)
- Categorical values in each of the non-temporal dimension
- Goal to find interesting patterns in the recent data as compared to the baseline data



ILLUSTRATION DATA: SINGLE DAY

- Imagine the data contains three dimensions:
 - Date
 - Age (C, A, S)
 - Disease (A, B, C, D)

	Disease			
	A	B	C	D
Age				
C				
A				
S				

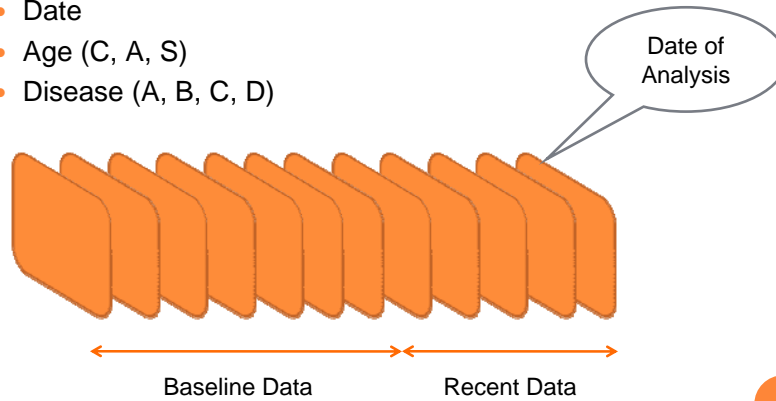
Number of records matching Age = {C} and Disease = {A} on a specific date



ILLUSTRATION DATA: MULTIPLE DAYS

○ Imagine the data contains three dimensions:

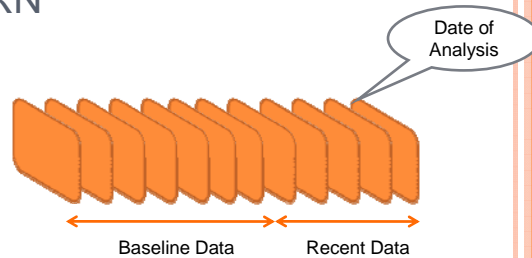
- Date
- Age (C, A, S)
- Disease (A, B, C, D)



DEFINING PATTERN

		Disease			
		A	B	C	D
Age	C				
	A				
	S				

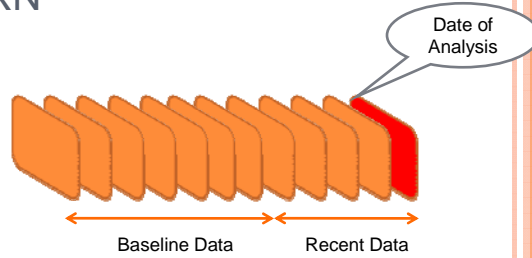
Pattern can take any subset of values in each of the non-temporal dimensions



Pattern must take a range of values in temporal dimension and must end on the date of analysis

DEFINING PATTERN

		Disease			
		A	B	C	D
Age	C		■		
	A				
	S				



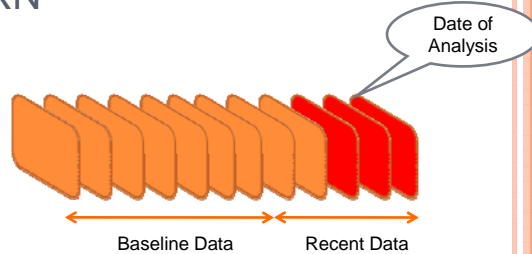
Age = {C} and Disease = {B} and Time = {T}

Conjunctive Pattern



DEFINING PATTERN

		Disease			
		A	B	C	D
Age	C			■	■
	A				
	S			■	■

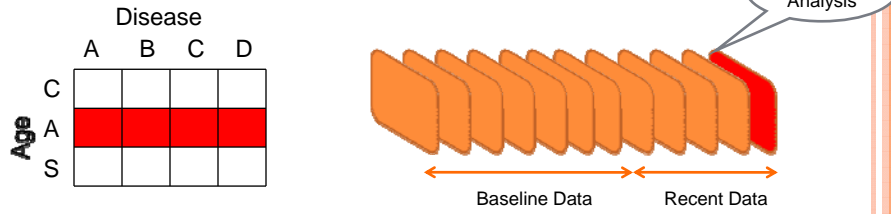


Age = {C, S} and Disease = {C, D} and Time = {T, T-1, T-2}

Disjunctive Pattern



DEFINING PATTERN



Age = {A} and T = {T}

All values affected in unspecified dimensions



PATTERN ANOMALOUS SCORE

	Recent	Baseline
Pattern		
~Pattern		

S_p = Fisher exact p-value



TALK OUTLINE

- Motivation
- Applications
- Multidimensional Data
- Disjunctive Pattern
- Pattern Scoring
- WSARE
- DAD
- Performance evaluation
- Conclusion



WHAT'S STRANGE ABOUT RECENT EVENTS: WSARE

- One of the few algorithms available for truly processing **multi-dimensional** data
- Works on dimensions with **categorical values**
 - age group, symptom, disease, gender, etc.
- Finds the single most interesting **conjunctive pattern** in recent data as compared to the baseline data



WSARE: ALGORITHM ILLUSTRATION

- Score all conjunctive patterns with one dimension and find the best scoring pattern

A = {C}
A = {A}
A = {S}
D = {A}
D = {B}
D = {C}
D = {D}



WSARE: ALGORITHM ILLUSTRATION

- Score all conjunctive patterns with one dimension and find the best scoring pattern
- Try to extend the best pattern by adding one extra dimension

A = {C}
A = {A}
A = {S}
D = {A}
D = {B}
D = {C}
D = {D}

A = {S}, D = {A}
A = {S}, D = {B}
A = {S}, D = {C}
A = {S}, D = {D}



WSARE: ALGORITHM ILLUSTRATION

- Score all conjunctive patterns with one dimension and find the best scoring pattern
- Try to extend the best pattern by adding one extra dimension
- When extending the pattern into higher dimensions, make sure the pattern is significant in all sub-components

$A = \{C\}$
 $A = \{A\}$
 $A = \{S\}$
 $D = \{A\}$
 $D = \{B\}$
 $D = \{C\}$
 $D = \{D\}$

$A = \{S\}, D = \{B\}$

	Recent	Baseline
$A = \{S\}, D = \{B\}$		
$A = \{S\}, D \neq \{B\}$		

	Recent	Baseline
$A = \{S\}, D = \{B\}$		
$A \neq \{S\}, D = \{B\}$		



DISJUNCTIVE ANOMALY DETECTION DAD

- Our goal is to find the **most anomalous disjunctive pattern** in **multidimensional** data
- Finding the highest scoring sub-matrix, in general, is an NP-hard problem. Hence, we propose a greedy algorithm (DAD).



DAD: ALGORITHM ILLUSTRATION

- Score all conjunctive patterns and store them in a queue sorted by ascending order of score.

A = {S} and D = {D} and T = {T}
A = {C}
A = {A}
A = {S}
D = {A}
...
A = {C} and D = {A}
..
A = {S} and D = {D}
A = {C} and T = {T}
...
...
A = {C} and D = {A} and T = {T}



DAD: ALGORITHM ILLUSTRATION

- Score all conjunctive patterns and store them in a queue sorted by ascending order of score.
- For every pattern in queue
 - try improving its score by adding at most one value to each dimension

A = {S} and D = {D} and T = {T}
A = {C}
A = {A}
A = {S}
D = {A}
...
A = {C} and D = {A}
..
A = {S} and D = {D}
A = {C} and T = {T}
...
...
A = {C} and D = {A} and T = {T}



DAD: ALGORITHM ILLUSTRATION

- Score all conjunctive patterns and store them in a queue sorted by ascending order of score.
- For every pattern in queue
 - try improving its score by adding at most one value to each dimension

A = {C, S} and D = {A}
A = {C, A} and D = {A}
A = {C} and D = {A, B}
A = {C} and D = {A, C}
A = {C} and D = {A, D}

A = {S} and D = {D} and T = {T}
A = {C}
A = {A}
A = {S}
D = {A}
...
A = {C} and D = {A}
..
A = {S} and D = {D}
A = {C} and T = {T}
...
...
A = {C} and D = {A} and T = {T}



DAD: ALGORITHM ILLUSTRATION

- Score all conjunctive patterns and store them in a queue sorted by ascending order of score.
- For every pattern in queue
 - try improving its score by adding at most one value to each dimension
 - Reinsert new pattern with better score in the queue

A = {S} and D = {D} and T = {T}
A = {C} and D = {A, B}
A = {C}
A = {A}
A = {S}
D = {A}
...
~~A = {C} and D = {A}~~
..
A = {S} and D = {D}
A = {C} and T = {T}
...
...
A = {C} and D = {A} and T = {T}



DAD: ALGORITHM ILLUSTRATION

- Score all conjunctive patterns and store them in a queue sorted by ascending order of score.
 - For every pattern in queue
 - try improving its score by
 - Report the best disjunctive anomalous pattern as the head of the queue

$A = \{S, C\}$ and $D = \{A\}$ and $T = \{T, T-1\}$
 $A = \{C\}$ and $D = \{A, B\}$
 $A = \{C\}$
 $A = \{A\}$
 $A = \{S\}$
 $D = \{A\}$
 ...
 $A = \{S\}$ and $D = \{D\}$
 $A = \{C\}$ and $T = \{T\}$
 ...

Better explanation of disease outbreak by reporting multiple values in each dimension

Constraint-based hill-climbing can handle spatial, temporal, dimension specific constraints during pattern search

ADDRESS PATTERN SCORE OVERFITTING

- Even in normal data, optimizing Fisher exact function scores larger patterns (more values along each dimension) better than smaller ones.
- To address this issue, we multiply the Fisher score of a pattern P by the number of other patterns in the data with size same as that of P
- This penalty is motivated by Bonferroni correction for multiple hypothesis testing.

TALK OUTLINE

- Motivation
- Applications
- Multidimensional Data
- Disjunctive Pattern
- Pattern Scoring
- WSARE
- DAD
- Performance evaluation
- Conclusion

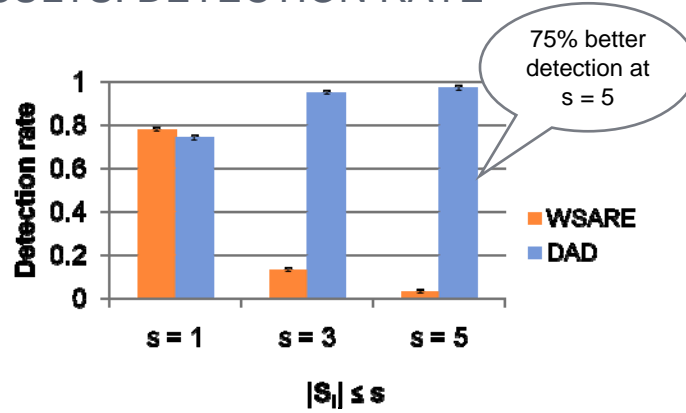


ALGORITHM EVALUATION

- Evaluation Data
 - Sri Lanka ministry of health care and nutrition data across 26 regions nationwide and 9 diseases.
 - Approximately 69,000 records over 30 months
 - We semi-synthetically added two more dimensions, namely, age group (10 values) and gender (2 values) for our analysis.
- Synthetic Patterns Injects
 - Ramp patterns with varying duration upto 7 days
 - Patterns with varying disjunctive complexity

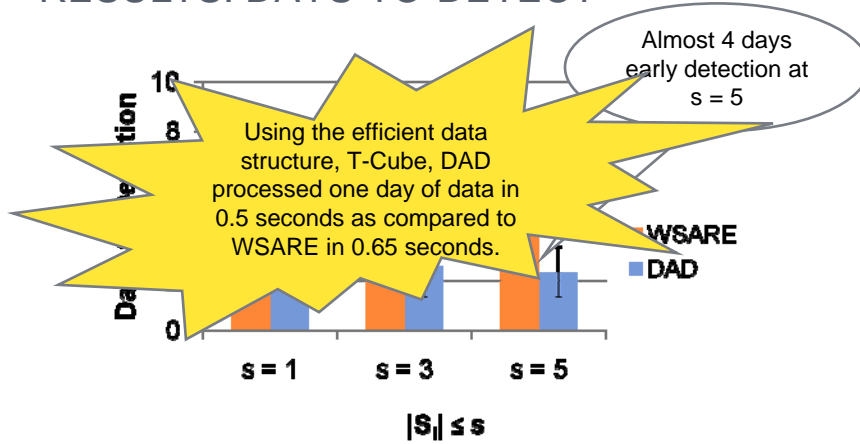


RESULTS: DETECTION RATE*



* average performance over 1,000 random ramp patterns each ranging from 1 to 7 days in duration and takes at most 's' values along each dimension

RESULTS: DAYS-TO-DETECT*



* average performance over 1,000 random ramp patterns each ranging from 1 to 7 days in duration and takes at most 's' values along each dimension

CONCLUSION

- Disease outbreaks with **disjunctive pattern explanations** are commonly found in the real world data.
- Finding such disjunctive patterns in the data is a challenging problem.
- We propose a greedy algorithm, **DAD**, to efficiently find realistic explanations of disease outbreaks.
- WSARE performs poorly in the presence of disjunctive anomalous patterns.



TAKE HOME MESSAGE

- For syndromic surveillance, we need to develop algorithms that efficiently mine multidimensional data and discover realistic disease outbreak explanations.
- The proposed algorithm, DAD, shows great promise towards achieving this goal.

