Provenance for Data Mining

Boris Glavic[1] Javed Siddique[2] Periklis Andritsos[3]
Renée J. Miller[2]

Illinois Institute of Technology[1]
DBGroup

University of Toronto[2]
Miller Lab

UofT:DB Group

University of Toronto[3]
iSchool

University of Toronto Faculty of Information

TaPP 2013, April 2, 2013

Outline

ILLINOIS INSTITUTE OF TECHNOLOGY

1. Motivation
2. Provenance for Data Mining
3. Frequent Itemset Provenance
4. Multidimensional Scaling Provenance
5. Conclusions

UofT:DB Group

UNIVERSITY OF TORONTO
FACULTY of INFORMATION

Goals

Extract useful information from data

Approach

1. Preprocessing
   - Cleaning
   - Feature Extraction
   - ...
2. Apply algorithms to extract information
   - Clustering, Frequent Itemset Mining, Classification, ...
   - Most approaches: Reduce size/Summarize data

Provenance for Data Mining?

Dilemmas

- Purpose of data mining necessitates summarization
  - “Needle in the haystack”
  - Loss of information
- Makes interpreting “raw” results harder
- User point of view: DM algorithm is black box

Provenance for Data Mining?

ILLINOIS INSTITUTE OF TECHNOLOGY

How to solve Dilemmas?

- Selective access to input data result is based on
  - Inputs to mining algorithm
  - Inputs to preprocessing
  - Contextual information
- Understand importance of inputs for results
  - Input data
  - Parameter settings
- Understand how data mining algorithm generates result from inputs
- Understand missing results

UofT:DB Group

UNIVERSITY OF TORONTO
FACULTY of INFORMATION

Provenance for Data Mining?

ILLINOIS INSTITUTE OF TECHNOLOGY

How to solve Dilemmas?

- Selective access to input data result is based on (Data Provenance+)
    - Inputs to mining algorithm
    - Inputs to preprocessing
    - Contextual information
- Understand importance of inputs for results (Responsibility)
    - Input data
    - Parameter settings
- Understand how data mining algorithm generates result from inputs (Process provenance)
- Understand missing results (Missing answers)

UofT:DB Group

UNIVERSITY OF TORONTO
FACULTY of INFORMATION

Related Work

Provenance

- Database provenance, Workflow provenance, Missing answers, Responsibility, . . .

Data Mining

- Enriching mining results with additional information
  - Contextual information for frequent itemsets[a]
- Visualization techniques[b]
- Determining effect of parameter settings/inputs on result
  - e.g., K-means cluster stability based on parameter settings[c]

[a] Q. Mei et al. "Generating semantic annotations for frequent patterns with context analysis". In: *SIGKDD*. 2006, pp. 337–346.

[b] D.A. Keim and H.P. Kriegel. "Visualization techniques for mining large databases: A comparison". In: *TKDE* 8.6 (1996), pp. 923–938.

[c] L.I. Kuncheva and D.P. Vetrov. "Evaluation of stability of k-means cluster ensembles with respect to random initialization". In: *TPAMI* 28.11 (2006), pp. 1798–1808.

Contributions

- Analyze requirements and use cases for data mining provenance (DMProv)
- Discuss applicability of existing approaches
- Outline challenges and sketch research directions
- Exemplify concepts on two concrete mining algorithms
  - Frequent Itemset Mining (FIM)
  - Multidimensional Scaling (MDS)

Outline

UofT:DB Group
UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

Fine-Grained Provenance

Why-Provenance

- Here Why-Provenance means all models based on influence
  - Subset of the input that caused output to appear in result
- "Caused by" modelled as
  - Sufficiency
  - Necessity
  - Preservation of Equivalence / Computability
  - Causality

Useful for Data Mining?

- Provenance concepts seem applicable to data mining
- Have to deal with large provenance size (summarization)
- Can abstract processing of classes of mining algorithms?
  - Do not reinvent provenance tracking for each algorithm!

Fine-Grained Provenance

ILLINOIS INSTITUTE OF TECHNOLOGY

Tracing Through Preprocessing Steps

- Track back data mining results to inputs of preprocessing
- ETL tools are used for preprocessing
- $\Rightarrow$ can use database or workflow provenance approaches?

UofT:DB Group

UNIVERSITY OF TORONTO FACULTY OF INFORMATION

Enriching Provenance with Context

ILLINOIS INSTITUTE OF TECHNOLOGY

Contextual Information as Provenance

- Mining algorithms often applied to a subset of available data
- Contextual Data: data related to the mining inputs
  - Automatic detection
  - User provided
- Contextual data often more usable and concise than provenance
- Which contextual data is of interest will differ
  - per use-case
  - maybe even per query
- $\Rightarrow$ Should support contextual data per provenance query/generation
  - Need flexible mechanism to select context (declarative?)

UofT:DB Group
UNIVERSITY OF TORONTO FACULTY OF INFORMATION

DMProv and Data Responsibility

ILLINOIS INSTITUTE OF TECHNOLOGY

Measuring Amount of Influence

- Single mining result influenced by large subset of input (all)
  - e.g., clustering
- Amount of influence differs significantly (**Responsibility**)

DB Responsibility Model

- **Causality**[a]
- **Counterfactual cause** i for output o
  - Removing i removes o from result
- **Actual cause** i for output o
  - **Contingency** C: Set of inputs to remove before i becomes CC for o
- **Responsibility**: $\frac{1}{\text{size of minimal contingency}}$

[a] A. Meliou et al. "Causality in databases". In: *IEEE Data Engineering Bulletin* 33.3 (2010), pp. 59–67; James Cheney. "Causality and the Semantics of Provenance". In: *DCM*. 2010, pp. 63–74.

DMProv and Data Responsibility

ILLINOIS INSTITUTE OF TECHNOLOGY

Applicability for Data Mining

- Reduce size of provenance
  - only return top-k responsible inputs in provenance
  - only return input with responsibility over threshold
- However: Output variables are not boolean

Solution Sketch

- Consider every input as a cause
- Consider every set of inputs as a contingency
- Measure amount of change
  - e.g., distance between cluster means
- Responsibility is sum of $\frac{1}{\text{size of contingency}} \cdot d(o, o')$ over all contingencies normalized by number of contingencies

UNIVERSITY OF TORONTO FACULTY OF INFORMATION

Data vs. Parameter Responsibility

Effect of Parameter Settings

- Mining results do depend on
  - Data
  - Parameter settings
- Define new responsibility type using both data and parameters
- Related Work: Robustness against parameter changes only
  - stability of clusterings[a]

[a] L.I. Kuncheva and D.P. Vetrov. "Evaluation of stability of k-means cluster ensembles with respect to random initialization". In: *TPAMI* 28.11 (2006), pp. 1798–1808.

- So far only data provenance
- Understand how mining algorithm combines inputs to produce an output
- Applicability of workflow and program analysis provenance techniques
    - Either too detailed or too coarse grained

Outline

1. Motivation
2. Provenance for Data Mining
3. **Frequent Itemset Provenance**
4. Multidimensional Scaling Provenance
5. Conclusions

Frequent Itemset Mining (FIM)

- One of the most prevalent mining tasks
- **Input**: set of transactions (sets of items)
  - Fixed domain $\mathbb{D}$
- **Output**: subsets of $\mathbb{D}$ (frequent itemsets)
  - appear in fraction larger σ (minimum support) of the transactions

FIM Example

ILLINOIS INSTITUTE OF TECHNOLOGY

Transaction

TID	Items	CID
1	{Coffee-mate, Coffee, Diaper, Beer}	1
2	{Diaper, Bread, Beer}	2
3	{Coffee-mate, Diaper, Coffee, Beer}	3
4	{Bread, Coffee}	4
5	{Coffee-mate, Coffee}	4
6	{Coffee-mate, Sugar}	4

FIM

FID	Frequent Items	Support
1	{Coffee}	4
2	{Coffee-mate}	4
3	{Diaper}	3
4	{Beer}	3
5	{Diaper, Beer}	3
6	{Coffee, Coffee-mate}	3

UofT:DB Group

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

FIM Example with Contextual Data

ILLINOIS INSTITUTE OF TECHNOLOGY

Transaction

TID	Items	CID
1	{Coffee-mate, Coffee, Diaper, Beer}	1
2	{Diaper, Bread, Beer}	2
3	{Coffee-mate, Diaper, Coffee, Beer}	3
4	{Bread, Coffee}	4
5	{Coffee-mate, Coffee}	4
6	{Coffee-mate, Sugar}	4

FIM

FID	Frequent Items	Support
1	{Coffee}	4
2	{Coffee-mate}	4
3	{Diaper}	3
4	{Beer}	3
5	{Diaper, Beer}	3
6	{Coffee, Coffee-mate}	3

Customer

CID	AgeGroup	Sex
1	20-40	m
2	20-40	m
3	20-40	m
4	50-60	f

UofT:DB Group

University of Toronto Faculty of Information

Why-Provenance for FIM

Intuition

- The transactions containing a frequent itemset I caused I to be frequent
- Define the Why-provenance as this set

Definition (Why-Provenance for FI)

- Given transaction base D, minimum support σ, itemset I
- $\mathcal{W}(I) = \{t \mid I \subseteq t \wedge t \in D\}$

UofT:DB Group

University of Toronto Faculty of Information

Transaction

TID	Items	CID
1	{Coffee-mate, Coffee, Diaper, Beer}	1
2	{Diaper, Bread, Beer}	2
3	{Coffee-mate, Diaper, Coffee, Beer}	3
4	{Bread, Coffee}	4
5	{Coffee-mate, Coffee}	4
6	{Coffee-mate, Sugar}	4

FIM

FID	Frequent Items	Support
1	{Coffee}	4
2	{Coffee-mate}	4
3	{Diaper}	3
4	{Beer}	3
5	{Diaper, Beer}	3
6	{Coffee, Coffee-mate}	3

Customer

CID	AgeGroup	Sex
1	20-40	m
2	20-40	m
3	20-40	m
4	50-60	f

Why-Provenance

FID	TIDs
1	{1,3,5,6}
2	{1,3,5,6}
3	{1,2,3}
4	{1,2,3}
5	{1,2,3}
6	{1,3,5,6}

FIM Provenance with Context

Customer

CID	AgeGroup	Sex
1	20-40	m
2	20-40	m
3	20-40	m
4	50-60	f

Why-Provenance

FID	TIDs
1	{1,3,5,6}
2	{1,3,5,6}
3	{1,2,3}
4	{1,2,3}
5	{1,2,3}
6	{1,3,5,6}

Example (Beer and Diaper)

- Beer and Diaper is frequent
- but why?

FIM Provenance with Context

Customer

CID	AgeGroup	Sex
1	20-40	m
2	20-40	m
3	20-40	m
4	50-60	f

Why-Provenance

FID	TIDs
1	{1,3,5,6}
2	{1,3,5,6}
3	{1,2,3}
4	{1,2,3}
5	{1,2,3}
6	{1,3,5,6}

Example (Beer and Diaper)

- Beer and Diaper is frequent
- but why?
- Why-provenance $\Rightarrow$ it appeared in this set of transactions

FIM Provenance with Context

Customer

CID	AgeGroup	Sex
1	20-40	m
2	20-40	m
3	20-40	m
4	50-60	f

Why-Provenance

FID	TIDs
1	{1,3,5,6}
2	{1,3,5,6}
3	{1,2,3}
4	{1,2,3}
5	{1,2,3}
6	{1,3,5,6}

Example (Beer and Diaper)

- Beer and Diaper is frequent
- but why?
- Why-provenance $\Rightarrow$ it appeared in this set of transactions
  - Not very useful!
  - Unfeasible if D is large

FIM Provenance with Context

ILLINOIS INSTITUTE OF TECHNOLOGY

Customer

CID	AgeGroup	Sex
1	20-40	m
2	20-40	m
3	20-40	m
4	50-60	f

Why-Provenance

FID	TIDs
1	{1,3,5,6}
2	{1,3,5,6}
3	{1,2,3}
4	{1,2,3}
5	{1,2,3}
6	{1,3,5,6}

Example (Beer and Diaper)

- Beer and Diaper is frequent
- but why?
- Why-provenance $\Rightarrow$ it appeared in this set of transactions
  - Not very useful!
  - Unfeasible if D is large
- Because male customers in age group $20 - 40$ bought it

FIM Provenance with Context

Customer

CID	AgeGroup	Sex
1	20-40	m
2	20-40	m
3	20-40	m
4	50-60	f

Why-Provenance

FID	TIDs
1	{1,3,5,6}
2	{1,3,5,6}
3	{1,2,3}
4	{1,2,3}
5	{1,2,3}
6	{1,3,5,6}

Example (Beer and Diaper)

- Beer and Diaper is frequent
- but why?
- Why-provenance $\Rightarrow$ it appeared in this set of transactions
  - Not very useful!
  - Unfeasible if D is large
- Because male customers in age group $20 - 40$ bought it
  - More useful and concise
  - Summarization of inputs in provenance using contextual information
  - Will need different context for different use cases

Preliminary Results

FIM Provenance

- Why-Provenance for FIM
- Declarative selection of context
- I-Provenance
  - Prefix compressed tree representation of provenance
  - Precise modelling of interdependencies of items in provenance within transactions
- Database-based provenance generation and querying

UofT:DB Group
UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

Outline

UofT:DB Group

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

Multidimensional Scaling (MDS)

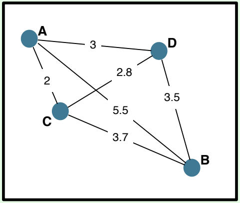
ILLINOIS INSTITUTE OF TECHNOLOGY

Approach

- **Input**: Set of observation with pair-wise similarities
- **Output**: Mapping into *n*-dim space that tries to preserve similarities
  - Optimization problem
- **Use-case Marketing**:
  - Customers rate products pairs according to similarity
  - MDS to generate layout (perceptual map) depicting similarity of products

Example

	A	B	C	D
A	-	-	-	-
B	2	-	-	-
C	2	4	-	-
D	3	3	3	-



Data vs. Parameter Responsibility

ILLINOIS INSTITUTE OF TECHNOLOGY

Problem

- If two items are close in the layout then
  - either they are similar
  - or because it minimized the fitness function
  - or some combination of both

Using Provenance

- Why-provenance
  - Show (difference to) original similarities for subset of the data
- Data vs. Parameter Responsibility
  - Influence of actual data properties
  - Parameter settings
  - Idiosyncrasies of the algorithm

UofT:DB Group

UNIVERSITY OF TORONTO FACULTY OF INFORMATION

Outline

UofT:DB Group

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

Challenges

Why-Provenance

- Common model that generalizes processing of large classes of mining algorithms
- Dealing with large (potentially overlapping) provenance

Context and Preprocessing

- Dynamic handling of contextual data
- Tracing through preprocessing steps

Responsibility

- Computational complexity
- How to model parameter vs. data responsibility?

Conclusions

Take Away Messages

- Data Mining is interesting and challenging application domain for provenance
- No previous work

Future Work

- Extend preliminary results on FIM
- Clustering (responsibility)
- MDS (parameter vs. data responsibility)

Questions?

- This is a vision paper
- so let's discuss!

UofT:DB Group

University of Toronto
Faculty of Information