On Statistical Schema Matching and Value Mapping

Anuj Jaiswal

Ph.D. Candidate,
College of Information Sciences and Technology,
The Pennsylvania State University

May 14, 2010

Outline I

Outline II

Future Work

Introduction

1. Driving forces for data integration and fusion
2. Data integration problem
3. Challenges
4. Our Approach

Driving forces for data integration and fusion

- Organizations evolving as *global* entities with distributed data, e.g., GM, Ford, Walmart
- Systems are characterized by a mix of *legacy* and *new* databases and applications
- Organizational *changes*
  1. *Growth* - size and diversity
  2. *Business re-engineering*
  3. Corporate mergers and acquisitions

Driving forces for data integration and fusion

- Organisations evolving as collections of distinct, autonomous departments with disconnected systems e.g. in financial services, NASA.
- Inter-operation between information sources over the web e.g. priceline.com and hotwire.com.

Data integration problem

- Data integration across heterogeneous sources involves two related subtasks:
  1. *Schema Matching*: "Structural alignment" of two schemas in two/ more sources
  2. *Value Mapping*: "Semantic resolution" of data across two/ more matched attributes
- Most research treats both problems independently

Schema Matching

- Reconciling structural alignment of data by matching schema attributes across information sources

Database A

Location	Price ($)	Agent-id
Atlanta, GA	360,000	12
York, PA	200,000	15

ID	Name	City	State
12	Agent Smith	Jackson	GA
15	Agent John	Mile Run	PA

Database B

Area	List-Price ($)	Agent Address	Agent Name
San Jose, CA	1,000,000	Ontario CA	Jane
Atlanta, GA	360,000	Jackson, GA	John Smith

Value Mapping

- Resolving semantic heterogeneity of data by mapping data value instances across matched schema attributes

Database A

Location	Price ($)	Agent-id
Atlanta, GA	360,000	12
York, PA	200,000	15

ID	Name	City	State
12	Agent Smith	Jackson	GA
15	Agent John	Mile Run	PA

Database B

Area	List-Price ($)	Agent Address	Agent Name
San Jose, CA	1,000,000	Ontario CA	Jane
Atlanta, GA	360,000	Jackson, GA	John Smith

Challenges

- Structural Conflicts
  - Naming Conflicts
- Attribute/Domain Conflicts
  - Data type conflicts
  - Measure and scale conflicts
  - Presence/absence of data values
  - Value semantics

Structural Conflicts

- Naming Conflicts
  - Multiple syntax/ description for data values
  - Synonyms: Customer vs. Client
  - Common Words: Name vs. Student Name
  - No semantic similarity (Opaque conditions)
    Four Wheel Drive Sedan vs. 4WD-S vs. Car Type 2
  - Similar names refer to semantically different attributes
    Name for Book domain vs. Name for University domain

Attribute/Domain Conflicts

- Data type (representation) conflicts
  - StudentID: 92115151 (Number vs. String)
  - StudentID vs. Student Name
- Measurement, scale, presision etc. conflicts
  - Measurement - Light years vs. Miles
  - Scale - Miles vs. Kilometers
  - Precision - Float vs. Double
  - Date formats - 12/02/1980 vs. 02/12/1980

Attribute/Domain Conflicts

- Presence/absence of data
  - Null entries present/absent
  - Matching attributes/ values missing
- Value semantics
  - Same items different values
  - Different items same values

Our Approach

- Automated technique which utilizes (*embeds*) value mappings to enhance schema matching [5]
  1. Designed for opaque conditions
  2. A global objective function to capture dissimilarity for fixed schema match and value mapping
     - attributes
     - attribute pairs
  3. Integrates both problems within a common frame work (minimization problem)
  4. Can tackle multiple feature spaces (categorical, continuous, mixed, transformed etc.)

Preliminaries

- Information Theory/ Probability basics
- Common related work
- Kang Naughton Mutual Information Methods
- Drawbacks

Information Theory/ Probability

- Entropy

$$H(X) = -\sum_{x \in \mathscr{X}} p(x) \log p(x) \tag{1}$$

- Joint Entropy

$$H(X, Y) = -\sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p(x, y) \log p(x, y) \tag{2}$$

- Conditional Entropy

$$H(X|Y) = -\sum_{x \in \mathscr{X}} \sum_{y \in \mathscr{Y}} p(x, y) \log p(x|y) \tag{3}$$

Information Theory/ Probability

- Mutual Information

$$MI(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (4)$$

- KL Divergence/ Relative Entropy

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad (5)$$

- Cross Entropy

$$h(p||q) = E_p[-\log q] = H(p) + D(p||q) \quad (6)$$

Information Theory/ Probability

- A likelihood function [3, 4] $L(\theta)$
  - probability or probability density for the occurrence of observations $(x_1, \ldots, x_n)$
  - given that the probability density $f(x;\theta)$ with parameter θ is known
- The likelihood function is defined as

$$L(\theta) = f(x_1;\theta)\ldots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \quad (7)$$

Related Work

- Linguistic/ Lexical Approaches
  - Uses lexical similarity between column names
  - E.g. Phone vs. HomePhone/ Office Phone
- Instance Level Approaches
  - Data instances of schemas are used to determine matches
  - E.g. Name vs. Student Name
- Hybrid Matchers
  - Combines multiple matching approaches to determine a set of ranked candidates
- Composite Matchers
  - Combines results to several matching techniques to determine match

Kang Naughton Technique

- Entropy based
  - Utilize statistics across single attributes
  - Align schemas by optimally reducing entropy across attributes
- Mutual Information (KN-MI) based
  - Utilize statistics based on attribute pairs
  - Build a dependency graph between attributes
  - Find schema match by utilizing a graph matching algorithm

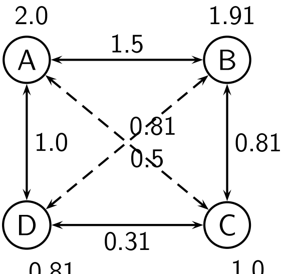
Kang Naughton Technique

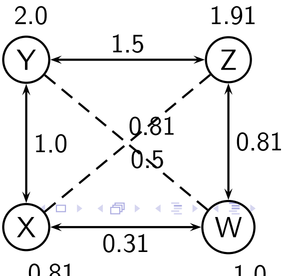
Database A

A	B	C	D
a1	b2	c1	d1
a2	b4	c2	d1
a3	b1	c1	d2
a4	b2	c1	d2

Database B

X	Y	Z	W
x1	y1	z2	w1
x1	y2	z4	w2
x2	y3	z1	w1
x2	y4	z2	w1





Drawbacks

- Scenario 1:
  - Entropy statistics do not align between schemas
  - Value cardinality cannot be used for decision making

Model (X)	P(X)	**Color (Y)**	P(Y)
XE	0.0472	White	0.0813
XL	0.0718	Blue	0.1077
XLE	0.1751	Red	0.1639
YE	0.2251	Silver	0.1702
YZ	0.2347	Black	0.1801
ZX	0.2461	Green	0.2968
H(X)	2.3938	**H(Y)**	2.4686

Table 1

A	P(A)	**B**	P(B)
X	0.0965	White	0.0733
XL	0.1010	Blue	0.0801
XLE	0.1075	Red	0.1011
GT	0.2044	Silver	0.2201
GE	0.2349	Black	0.2418
GL	0.2557	Green	0.2836
H(A)	2.4677	**H(B)**	2.3939

Table 2

Drawbacks

- Scenario 2:
  - Entropy statistics are similar within schema
  - Value cardinality cannot be used for decision making

Model (X)	P(X)	**Color (Y)**	P(Y)
XE	0.0472	White	0.0813
XL	0.0718	Blue	0.1077
XLE	0.1751	Red	0.1639
YE	0.2251	Silver	0.1702
YZ	0.2347	Black	0.1801
ZX	0.2461	Green	0.2968
H(X)	2.3938	**H(Y)**	2.4686

Table 1

A	P(A)	**B**	P(B)
X	0.0244	White	0.1603
XL	0.0570	Blue	0.1653
XLE	0.1796	Red	0.1846
GT	0.2047	Silver	0.2343
GE	0.2536	Black	0.2554
GL	0.2807		
H(A)	2.2962	**H(B)**	2.2962

Table 2

Drawbacks

- Scenarios may apply to
  - Entropy
  - Mutual Information
  - Conditional Entropy

The Framework

- Capitalize on the value mapping dimension
  1. Utilize frequency of occurence of an attribute's value set
  2. Categorical Attributes
     - Probability mass function(pmf) of an attribute's value set is more distinctive than entropy
  3. Continuous Attributes
     - Probability density function(pdf) of an attribute's value set is more distinctive than the pmf/ entropy
  4. Mixed Attributes
     - Utilize pmf's and pdf's to model such schemas
  5. Matching pmf's and pdf's across schemas should be more reliable
     - Second-order statistics for finer matching

Modelling Dissimilarity of Match

- Model dissimilarity score
  - A global objective function which provides a confidence score for a fixed schema mapping and value mapping
  - Uses a dissimilarity metric based on
    - Euclidean distance
    - Log-Likelihood
    - Relative Entropy
- Two dimensional minimization of global objective
  1. First dimension - Schema Matching
  2. Second dimension - *Embedded* Value Mapping

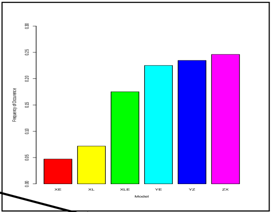
Modelling Dissimilarity of Match

- First-order dissimilarity metric models
  - Measures how well the probability mass functions (and/or) probability distribution functions align for fixed schema matching
  - Uses first-order statistics of attributes
- Second-order dissimilarity metric models
  - Measures how well the joint probability mass functions (and/or) joint probability distribution functions align for fixed schema matching
  - Uses second-order statistics between attribute pairs

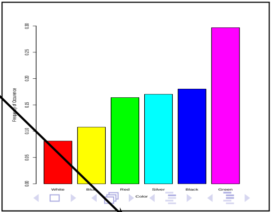
Categorical Feature Spaces

Model (X)	P(X)	**Color (Y)**	P(Y)
XE	0.0472	White	0.0813
XL	0.0718	Blue	0.1077
XLE	0.1751	Red	0.1639
YE	0.2251	Silver	0.1702
YZ	0.2347	Black	0.1801
ZX	0.2461	Green	0.2968
H(X)	2.3938	**H(Y)**	2.4686

Table 1



PMF of Model



PMF of Color

First-order Dissimilarity Models

- Euclidean squared distance based model

$$\mathrm{D}_{AB}^{EU} = \sum_{i=1}^{N_{attr_1}} \sum_{i'=1}^{N_{attr_2}} \delta(i - m_a(i')) \sum_{j=1}^{N_{Values(i)}} \left[p_i(j) - \left\{ \sum_{j'=1}^{N_{Values(i')}} \delta(j - M_v^{(i,i')}(j')) p_{i'}(j') \right\} \right]^2 \tag{8}$$

- Relative Entropy based model

$$\mathrm{D}_{AB}^{CE} = \sum_{i=1}^{N_{attr_1}} \sum_{i'=1}^{N_{attr_2}} \delta(i - m_a(i')) \sum_{j=1}^{N_{Values(i)}} \left[p_i(j) \times log \frac{p_i(j)}{[\sum_{j'=1}^{N_{Values(i')}} \delta(j - M_v^{(i,i')}(j')) p_{i'}(j')]} \right] \tag{9}$$

First-order Models: Issues

- Advantages of Euclidean-distance vs. Relative Entropy
  - Relative Entropy requires $N_{Values(i)} = N_{Values(i')}$
  - Value alphabet set must be same for matching attributes
  - Requires Attribute Alphabet reconstruction
    - Introduce extra symbols and assign them small probability followed by normalization
    - Reduce size of the larger alphabet by deleting the smallest probability values followed by normalization
    - Euclidean-distance is free from such requirements
  - Relative Entropy is not symmetric
  - Euclidean-distance is computationally cheaper

Second-order Dissimilarity Models

- Euclidean-squared distance based model

$$\mathrm{D}_{AB}^{PEU} = \sum_{i=1}^{N_{attr_1}} \sum_{\substack{j=1 \\ i \neq j}}^{N_{attr_1}} \left[\sum_{i'=1}^{N_{attr_2}} \delta(i - m_a(i')) \sum_{\substack{j'=1 \\ i' \neq j'}}^{N_{attr_2}} \delta(j - m_a(j')) \right.$$
$$\left. \left[\sum_{k=1}^{N_{V(i)}} \sum_{l=1}^{N_{V(j)}} \left[p_{ij}(k,l) - \left\{ \sum_{k'=1}^{N_{V(i')}} \sum_{l'=1}^{N_{V(j')}} \delta(k' - M_v^{(i,i')}(k)) \delta(l' - M_v^{(j,j')}(l)) p_{i'j'}(k',l') \right\} \right]^2 \right] \right] \tag{10}$$

Matching and Mapping Strategy

- Our global objective is now two-dimensional minimization of the dissimilarity objective

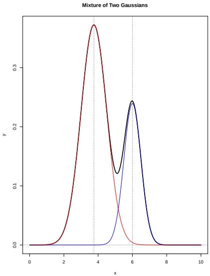
$$\mathrm{D}_{Overall} = \min_{x \in S} \left\{ \min_{y \in V} [D^{M}_{AB}] \right\} \tag{11}$$

Schema Matching Dimension

Value Mapping Dimension

Continuous Feature Spaces

Price ($) (X)	Cost
2	10
3	11
...	...
4	1
5	1

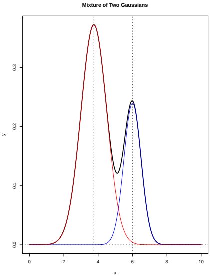


$$\alpha_1 = 0.7 \ \mu_1 = 3.75 \ \sigma_1 = 0.75$$
$$\alpha_2 = 0.3 \ \mu_2 = 6.00 \ \sigma_2 = 0.50$$

$$f(x_i;\theta_i) = \sum_{n=1}^{K_i} \alpha_n^i \phi(x_i | \mu_n^i, \sigma_n^i)$$
$$\sum \alpha_n^i = 1$$
$$0 \leq \alpha_n^i \leq 1$$

Continuous Feature Spaces

Price ($) (X)	Cost
2	10
3	11
...	...
4	1
5	1



$\alpha_1 = 0.7$ $\mu_1 = 3.75$ $\sigma_1 = 0.75$

$\alpha_2 = 0.3$ $\mu_2 = 6.00$ $\sigma_2 = 0.50$

$$f(x_i;\theta_i) = \sum_{n=1}^{K_i} \alpha_n^i \phi(x_i|\mu_n^i, \sigma_n^i)$$

$$\sum \alpha_n^i = 1$$

$$0 \leq \alpha_n^i \leq 1$$

First-order Dissimilarity Models

- Log-Likelihood based model

$$D_{AB}^{LL} = \sum_{i=1}^{N_{C_1}} \sum_{j=1}^{N_{C_2}} \delta(i - m_a(j)) \left[\sum_{a=1}^{N_{R_1}} log \left[\sum_{n=1}^{K_{yj}} \beta_{yn}^{j} N(x_{ia} | \mu_{yn}^{j}, \sigma_{yn}^{j}) \right] + \sum_{a=1}^{N_{R_2}} log \left[\sum_{n=1}^{K_{xi}} \alpha_{xn}^{i} N(y_{ja} | \mu_{xn}^{i}, \sigma_{xn}^{i}) \right] \right] \tag{12}$$

- Euclidean-distance based models
  - Ordered Component Model

$$D_{AB}^{EU} = \sum_{i=1}^{N_{C_1}} \sum_{j=1}^{N_{C_2}} \delta(i - m_a(j)) \left[\sum_{k=1}^{K_{xi}} \sum_{k'=1}^{K_{yj}} (\alpha_{xk}^{i} - \delta(k - m_c^{(i,j)}(k')) \beta_{yk'}^{j})^2 \right] \tag{13}$$

  - Model Estimation Model

$$D_{AB}^{ME} = \sum_{i=1}^{N_{C_1}} \sum_{j=1}^{N_{C_2}} \delta(i - m_a(j)) \left[\sum_{k=1}^{K_{xi}} (\alpha_{xk}^{i} - \hat{\alpha}_{yk}^{i})^2 + \sum_{k'=1}^{K_{yj}} (\beta_{yk'}^{j} - \hat{\beta}_{xk'}^{j})^2 \right] \tag{14}$$

Second-order Dissimilarity Models

- Log-Likelihood based model

$$D_{AB}^{LL} = \sum_{i=1}^{N_{C_1}} \sum_{\substack{j=1 \\ i \neq j}}^{N_{C_1}} \left[\sum_{i'=1}^{N_{C_2}} \sum_{\substack{j'=1 \\ i' \neq j'}}^{N_{C_2}} \delta(i - m_a(i'))\delta(j - m_a(j')) \left[\sum_{a=1}^{N_{R_1}} log \left[\sum_{n=1}^{K_{yi'j'}} \beta_{yn}^{i'j'} N(x_{ija} | \mu_{yn}^{i'j'}, \sigma_{yn}^{i'j'}) \right] \right. \right.$$
$$\left. \left. + \sum_{a=1}^{N_{R_2}} log \left[\sum_{n=1}^{K_{xij}} \alpha_{xn}^{ij} N(y_{i'j'a} | \mu_{xn}^{ij}, \sigma_{xn}^{ij}) \right] \right] \right] \quad (15)$$

- Euclidean-distance based models
  - Ordered Component Model

$$D_{AB}^{EU} = \sum_{i=1}^{N_{C_1}} \sum_{\substack{j=1 \\ i \neq j}}^{N_{C_1}} \left[\sum_{i'=1}^{N_{C_2}} \sum_{\substack{j'=1 \\ i' \neq j'}}^{N_{C_2}} \delta(i - m_a(i'))\delta(j - m_a(j')) \left[\sum_{k=1}^{K_{xij}} \sum_{k'=1}^{K_{yi'j'}} (\alpha_{xk}^{ij} - \delta(k - m_c^{(ij,i'j')}(k'))\beta_{yk'}^{i'j'})^2 \right] \right] \quad (16)$$

Second-order Dissimilarity Models

- Euclidean-distance based models
  - Model Estimation

$$D_{AB}^{ME} = \sum_{i=1}^{N_{C_1}} \sum_{\substack{j=1 \\ i \neq j}}^{N_{C_1}} \left[\sum_{i'=1}^{N_{C_2}} \sum_{\substack{j'=1 \\ i' \neq j'}}^{N_{C_2}} \delta(i - m_a(i'))\delta(j - m_a(j')) \left[\sum_{k=1}^{K_{xij}} (\alpha_{xk}^{ij} - \hat{\alpha}_{yk}^{ij})^2 + \sum_{k'=1}^{K_{yi'j'}} (\beta_{yk'}^{i'j'} - \hat{\beta}_{xk'}^{i'j'})^2 \right] \right] \quad (17)$$

Matching and Mapping Strategy

- Our global objective is now one-dimensional minimization/ maximization of the dissimilarity objective

$$\mathrm{D}_{Overall} = \min_{x \in S} [D_{AB}^{MM}] \tag{18}$$

Or,

$$\mathrm{D}_{Overall} = \max_{x \in S} [D_{AB}^{LL}] \tag{19}$$

Schema Matching Dimension

Embedded Value Mapping Dimension

Matching and Mapping Strategy

- Our global objective is now one-dimensional minimization/ maximization of the dissimilarity objective

$$\mathrm{D}_{Overall} = \min_{x \in S} [D_{AB}^{MM}] \tag{18}$$

Or,

$$\mathrm{D}_{Overall} = \max_{x \in S} [D_{AB}^{LL}] \tag{19}$$

Schema Matching Dimension

Embedded Value Mapping Dimension

Matching and Mapping Strategy

- Our global objective is now one-dimensional minimization/ maximization of the dissimilarity objective

$$\mathrm{D}_{Overall} = \min_{x \in S} [D_{AB}^{MM}] \tag{18}$$

Or,

$$\mathrm{D}_{Overall} = \max_{x \in S} [D_{AB}^{LL}] \tag{19}$$

Schema Matching Dimension

Embedded Value Mapping Dimension

Mixed Feature Spaces

- We can seperate this problem into two sub-problems
  - Categorical Attributes and use discrete dissimilarity measures
  - Continuous Attributes and use continuous dissimilarity measures
  - Merge results to arrive at final schema match

OR,

- Define a global objective which can tackle such types of tables simultaneously
  - Attributes are defined by either probability mass functions or probability distribution functions
  - Global objective is sum of compatible dissimilarity measures for the different feature spaces

First-order Dissimilarity Models

- Euclidean squared distance based models

$$\mathrm{D}_{AB}^{EU} = \sum_{i=1}^{N_{D_1}} \sum_{i'=1}^{N_{D_2}} \delta(i - m_a(i')) \sum_{j=1}^{N_{Values(i)}} \left[p_i(j) - \left\{ \sum_{j'=1}^{N_{Values(i)}} \delta(j - M_v^{(i,i')}(j')) p_{i'}(j') \right\} \right]^2 + \sum_{j=1}^{N_{C_1}} \sum_{j'=1}^{N_{C_2}} \delta(j - m_a(j')) \left[\sum_{k=1}^{K_{xj}} \sum_{k'=1}^{K_{yj'}} (\alpha_{xk}^{j} - \delta(k - m_c^{(j,j')}(k')) \beta_{yk'}^{j'})^2 \right] \quad (20)$$

$$\mathrm{D}_{AB}^{EU} = \sum_{i=1}^{N_{D_1}} \sum_{i'=1}^{N_{D_2}} \delta(i - m_a(i')) \sum_{j=1}^{N_{Values(i)}} \left[p_i(j) - \left\{ \sum_{j'=1}^{N_{Values(i)}} \delta(j - M_v^{(i,i')}(j')) p_{i'}(j') \right\} \right]^2 + \sum_{j=1}^{N_{C_1}} \sum_{j'=1}^{N_{C_2}} \delta(j - m_a(j')) \left[\sum_{k=1}^{K_{xj}} (\alpha_{xk}^{j} - \hat{\alpha}_{yk}^{j})^2 + \sum_{k'=1}^{K_{yj'}} (\beta_{yk'}^{j'} - \hat{\beta}_{xk'}^{j'})^2 \right] \quad (21)$$

First-order Dissimilarity Models

- Log-Likelihood based model

$$D_{AB}^{LL} = \sum_{i=1}^{N_{D_1}} \sum_{i'=1}^{N_{D_2}} \delta(i - m_a(i'))\Big[\sum_{j=1}^{N_{V(i)}} \sum_{j'=1}^{N_{V(i')}} \delta(j - M_v^{(i,i')}(j'))n_i(j)p_{i'}(j') + \sum_{j=1}^{N_{V(i)}} \sum_{j'=1}^{N_{V(i')}} \delta(j - M_v^{(i,i')}(j'))n_{i'}(j')p_i(j)\Big]$$
$$+ \sum_{i=1}^{N_{C_1}} \sum_{j=1}^{N_{C_2}} \delta(i - m_a(j))\Big[\sum_{a=1}^{N_{R_1}} log\Big[\sum_{n=1}^{K_{yj}} \beta_{yn}^{j} N(x_{ia}|\mu_{yn}^{j}, \sigma_{yn}^{j})\Big] + \sum_{a=1}^{N_{R_2}} log\Big[\sum_{n=1}^{K_{xi}} \alpha_{xn}^{i} N(y_{ja}|\mu_{xn}^{i}, \sigma_{xn}^{i})\Big]\Big] \quad (22)$$

First-order Models: Issues

- Advantages of Euclidean-distance vs. Log-Likelihood
  - Log-Likelihood requires cardinality of matching pmf's to be same
  - Requires Attribute Alphabet reconstruction
    - Introduce extra symbols and assign them small probability followed by normalization
    - Use the pmf of larger cardinality in measuring the dissimilarity
    - Euclidean-distance is free from such requirements
  - Euclidean-distance is computationally cheaper

Matching and Mapping Strategy

- Our global objective is now one-dimensional minimization/ maximization of the dissimilarity objective

$$\mathrm{D}_{Overall} = \min_{x \in S} [D_{AB}^{MM}] \tag{23}$$

Or,

$$\mathrm{D}_{Overall} = \max_{x \in S} [D_{AB}^{LL}] \tag{24}$$

Schema Matching Dimension

Embedded Value Mapping Dimension

Matching and Mapping Strategy

- Our global objective is now one-dimensional minimization/ maximization of the dissimilarity objective

$$\mathrm{D}_{Overall} = \min_{x \in S} [D_{AB}^{MM}] \tag{23}$$

Or,

$$\mathrm{D}_{Overall} = \max_{x \in S} [D_{AB}^{LL}] \tag{24}$$

Schema Matching Dimension

Embedded Value Mapping Dimension

Matching and Mapping Strategy

- Our global objective is now one-dimensional minimization/ maximization of the dissimilarity objective

$$\mathrm{D}_{Overall} = \min_{x \in S} [D_{AB}^{MM}] \tag{23}$$

Or,

$$\mathrm{D}_{Overall} = \max_{x \in S} [D_{AB}^{LL}] \tag{24}$$

Schema Matching Dimension

Embedded Value Mapping Dimension

Experiments and Results

- Data Analyzer and Modelling
  - Loads data tables
  - Performs density estimation for continuous/ mixed datasets
  - Learns *latent* categorical variables for continuous attributes
    - Use maximum a posteriori (MAP) estimate to generate pmf's
- Schema Matching with *Embedded* Value Mapping algorithms take over
  - Measure precision of schema match over *50* iterations
  - Each iteration corresponds to a *different* schema matching problem

Implementation

- Software
  - GCC 3.4.5, GSL Scientific Library
  - R [7], Mclust [1]
- Hardware
  - Cyberstar cluster
  - Inti cluster

Datasets

- We used real world datasets from the Census Bureau
  - 1990 Public-Use Microdata Samples (PUMS) 5% Sample for states of California, New York, Texas
  - Texas Dataset
    - 231 Attributes
    - 935K records
  - California Dataset
    - 231 Attributes
    - 1500K records
  - New York Dataset
    - 231 Attributes
    - 960K records

Dataset Creation

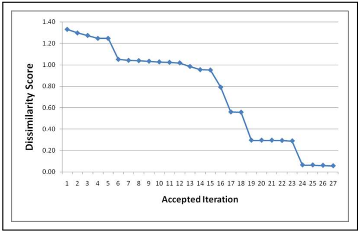
- We created three datasets for testing out algorithms
  - Categorical Dataset
    - Selected 30 categorical columns (Value Alphabet ≤ 50)
    - Some columns were discretized using coding from Meek et. al. [6]
    - Example attributes: "Race", "Age" etc.
  - Continuous Dataset
    - Selected 26 pure continuous columns (Value Alphabet ≥ 90)
    - Example attributes: "Water Cost", "Age" etc.
  - Mixed Dataset
    - Selected 15 pure continuous columns (Value Alphabet ≥ 90)
    - Selected 15 categorical columns (Value Alphabet ≤ 50)

Categorical Dataset

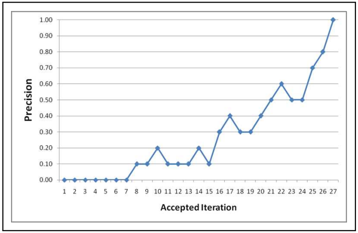
- Second-order Model
  - Each attribute pair is modelled using a joint probability mass function
- First-order Model
  - Each attribute is modelled using a probability mass function
- Euclidean-squared dissimilarity metric model used
  - Free from Value Alphabet Reconstruction issues

Monotonic Nature

- Evolution of dissimilarity score as a function of *accepted* iteration
  - Schema matching on 10 attributes across NY and CA datasets
  - Accepted iteration obtained by 2-opt switch
  - and decreases the objective function

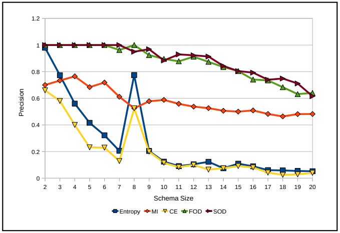


Euclidean-squared Dissimilarity Score vs. Accepted Iteration

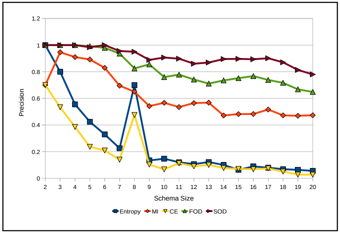


Precision vs. Accepted Iteration

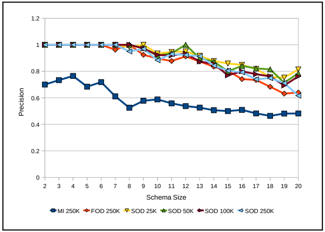
One-to-One Schema Matching



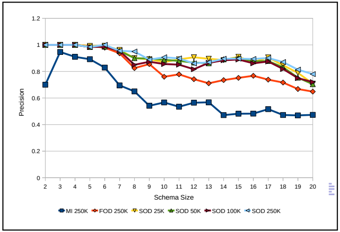
Average precision for CA vs. NY dataset



Average precision for CA vs. TX dataset



Effect of rows on precision for CA vs. NY dataset



Effect of rows on precision for CA vs. TX dataset

Continuous Dataset

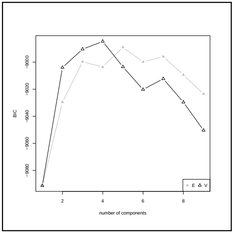
- Mixture model component size for each attribute is estimated using Bayesian Information Criterion (BIC)
- BIC is defined as

$$BIC = -2 \log L(x;\theta) + k \log n \tag{25}$$

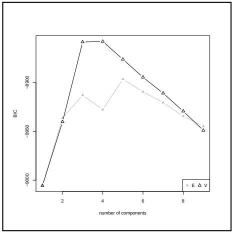
- Component size is varied as

$$1 \leq k \leq 20$$

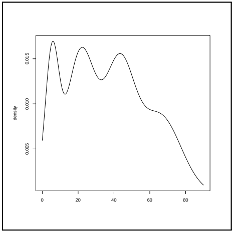
Continuous Attribute Modelling



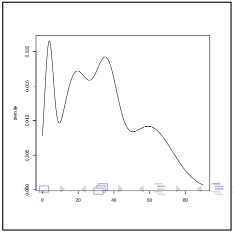
BIC values for varying component size for New York dataset



BIC values for varying component size for California Dataset



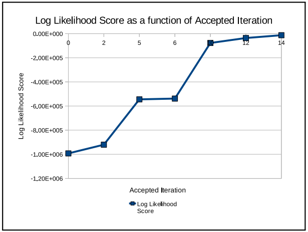
Estimated Density for “Age” using 4 components for New York dataset



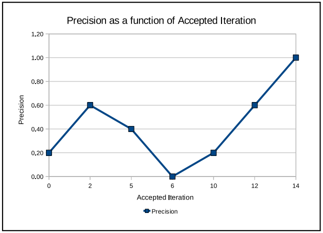
Estimated Density for “Age” using 4 components for California Dataset

Monotonic Nature

- Evolution of dissimilarity score as a function of *accepted* iteration
  - Schema matching on 10 attributes across NY and CA datasets
  - Accepted iteration obtained by 2-opt switch
  - and increases the objective function

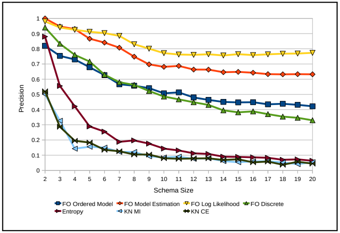


Log-Likelihood Dissimilarity Score vs. Accepted Iteration

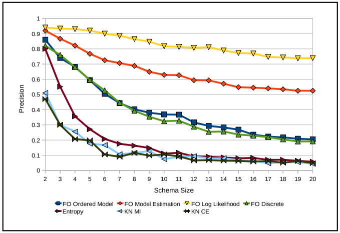


Precision vs. Accepted Iteration

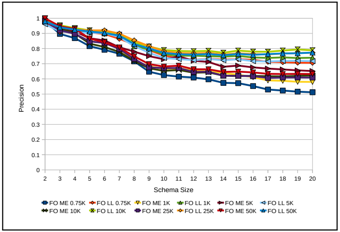
One-to-One Schema Matching



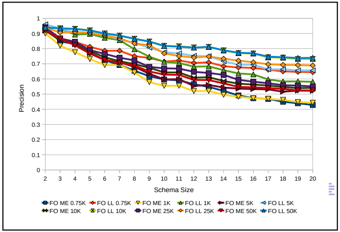
Average precision for CA vs. NY dataset



Average precision for CA vs. TX dataset

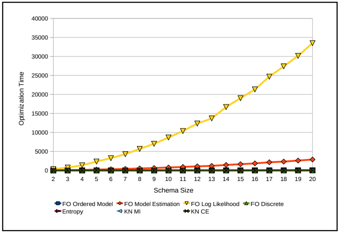


Effect of rows on precision for CA vs. NY dataset

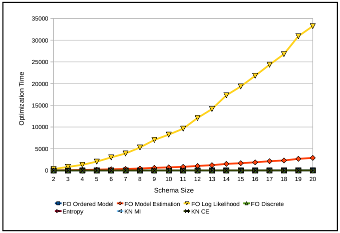


Effect of rows on precision for CA vs. TX dataset

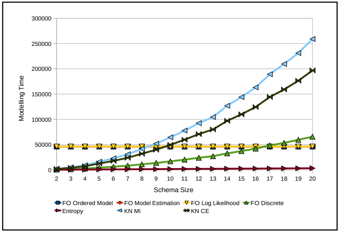
Computation Time



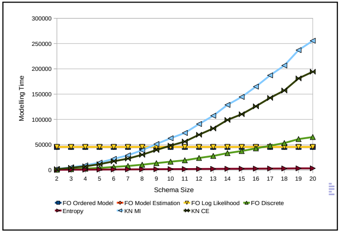
Computation Time for optimization for CA vs. NY dataset



Computation Time for optimization for CA vs. TX dataset



Computation Time for modelling for CA vs. NY dataset



Computation Time for modelling for CA vs. TX dataset

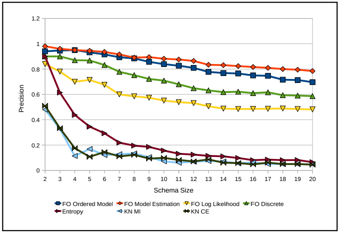
Mixed Dataset

- Each continous attribute is modelled by a mixture of normals
  - Component size is estimated using Bayesian Information Criterion (BIC)
  - Component size is varied as

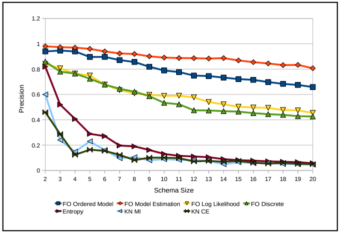
$$1 \leq k \leq 20$$

- Each categorical attribute is modelled by a probability mass function

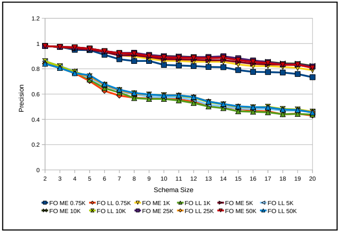
One-to-One Schema Matching



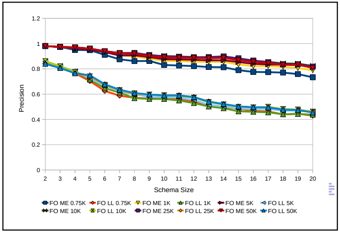
Average precision for CA vs. NY dataset



Average precision for CA vs. TX dataset

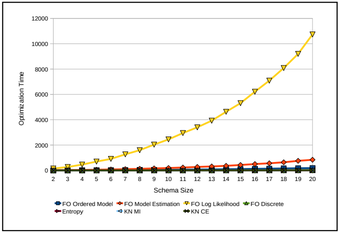


Effect of rows on precision for CA vs. NY dataset

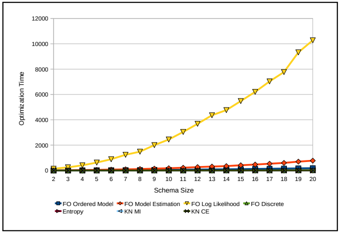


Effect of rows on precision for CA vs. TX dataset

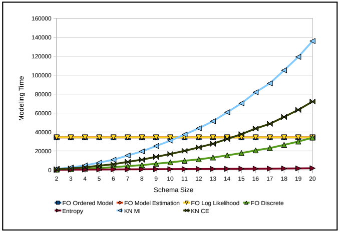
Computation Time



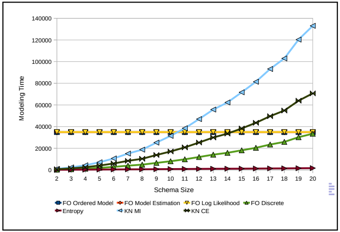
Computation Time for optimization for CA vs. NY dataset



Computation Time for optimization for CA vs. TX dataset



Computation Time for modelling for CA vs. NY dataset



Computation Time for modelling for CA vs. TX dataset

Experimental Summary

- Categorical Feature Spaces
  - First-order [5] and Second-order dissimilarity metrics evaluated
  - Random Initialization
  - Ground-truth Initialization
  - Fixed Initialization
- Mixed and Continuous Feature Spaces
  - First-order dissimilarity metrics evaluated *(To be submitted to pVLDB)*
  - Random Initialization
  - Ground-truth Intialization
  - Fixed Initialization

Research Contributions

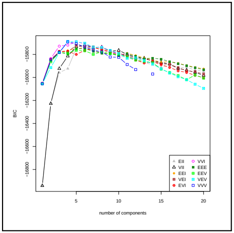
- No previous work which tackles both schema matching and value mapping at the same time.
- Extended *embedded* value mapping technique to tackle
  - Continuous attributes
  - Mixed attributes

Future Work

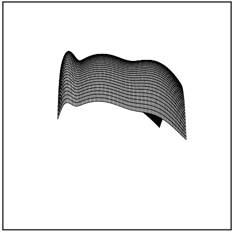
- Evaluating second-order models
- Transformed Feature Spaces
- Schema Matching Criterion

Second-order Models

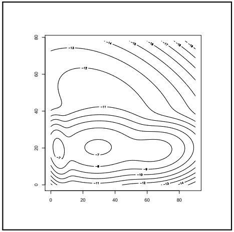
- Continuous and Mixed Feature Spaces
  - Experimental evaluation of second-order models
  - Finer matching should be achieved



Bayesian Information Criterion scores to determine component size for attribute pair "Age" and "Weight" for the California dataset. Best BIC score is obtained for 4 components.



Density estimate perspective for attribute pair "Age" and "Weight" for California Dataset



Density estimate contour for attribute pair "Age" and "Weight" for California Dataset

Transformed Feature Spaces

- Continuous Dissimilarity Metric Models
  - Assume no transformation necessary between two matching attributes
- However, real world datasets might not have one-to-one correspondence between data values
  - "Temperature" might be measured in "Celcius" or "Fahrenheit"
  - "Distance" may be measured in "Kilometers" or "Miles"
  - Two similar attributes would match only when a *transformation* operator is applied

Transformed Feature Spaces

- Continous dissimilarity models can be extended to tackle such situation
  - We consider **affine** transformations

$$x \mapsto Ax + b \tag{26}$$

- Consider set of pre-defined transformations

$$\mathbf{T} = [T_1, \ldots, T_{N_T}]^T \tag{27}$$

First-order Dissimilarity Model

- Log-Likelihood based model

$$D^{LL}_{AB} = \sum_{i=1}^{N_{C_1}} \sum_{j=1}^{N_{C_2}} \delta(i - m_a(j)) \Big[\sum_{a=1}^{N_{R_1}} log \Big[\sum_{n=1}^{K_{yj}} \beta^j_{yn} N(T_z[x_{ia}] | \mu^j_{yn}, \sigma^j_{yn}) \Big] + \sum_{a=1}^{N_{R_2}} log \Big[\sum_{n=1}^{K_{xi}} \alpha^i_{xn} N(T_z^{-1}[y_{ja}] | \mu^i_{xn}, \sigma^i_{xn}) \Big] \Big] \tag{28}$$

- Euclidean-distance based models
  - Model Estimation Model

$$D^{ME}_{AB} = \sum_{i=1}^{N_{C_1}} \sum_{j=1}^{N_{C_2}} \delta(i - m_a(j)) \Big[\sum_{k=1}^{K_{xi}} (\alpha^i_{xk} - \hat{\alpha}^{iT_z^{-1}}_{yk})^2 + \sum_{k'=1}^{K_{yj}} (\beta^j_{yk'} - \hat{\beta}^{jT_z}_{xk'})^2 \Big] \tag{29}$$

Matching and Mapping Strategy

- Our global objective is now two-dimensional minimization/maximization of the dissimilarity objective

$$\mathrm{D}_{Overall} = \min_{\substack{x \in S \\ T_z \in T}} [D^{ME}_{ABTz}] \tag{30}$$

Or,

$$\mathrm{D}_{Overall} = \max_{\substack{X \in S \\ T_z \in T}} [D^{LL}_{ABTz}] \tag{31}$$

Embedded Value Mapping Dimension

Schema Matching Dimension

Transformation Operator Dimension

Experimental Validation

- Datasets
  - Create a transformed dataset from Census Continuous dataset
- Extend to Second-order model for dealing with transformed spaces
- Integrate transformation learning within framework
  - Regression analysis techniques for learning operator
    - Ordinary least squares (OLS)
  - Complexity of problem will grow as incorrect attributes might now match better

Schema Matching Criterion

- One-to-One schema matching
  - Each attribute in *schema 1* has unique matching attribute in *schema 2* and vice versa
  - Equal number of attributes in both schemas
  - Our methods deal with this type of schema matching
- Onto or Subset schema matching
  - Each attribute in *schema 1* has unique matching attribute in *schema 2*
  - Each attribute in *schema 2* has unique matching attribute in *schema 1* or remains unmatched
  - Unequal number of attributes across schemas
  - Can be posed as an assignment problem

Schema Matching Criterion

- Partial schema matching
  - Each attribute in *schema 1* has a unique matching attribute in *schema 2* or remains unmatched, and vice versa
  - Unknown number of attributes match
  - Investigate methods for proposing confident matches
    - Regularization: Adapting the BIC measure for model selection
    - Hypothesis testing

Schema Matching Criterion

- Many-to-many schema matching
    - Each attribute in *schema 1* has a set of matching attributes in *schema 2* or remains unmatched, and vice versa
    - Most difficult and general problem
    - Solving this problem automatically may not be feasible (user input maybe required)
    - Investigate methods for proposing top-K [2] candidate attributes that may be matched

Questions?

Backup Slides

Categorical Schema Matching Algorithm

Algorithm 1 Overview of our approach

Input: Schemas, *T1* and *T2*
Output: Schema Match *s*, *tmpS*
Value Mapping *v*, *tmpV*, *v1*
sMatchScore $\leftarrow \infty$ { Best Dissimilarity Score}
sSpace $\leftarrow$ getSchemaMatchSearchSpace()
while *sSpace* is not empty **do**
  tmpS $\leftarrow$ getNextSchemaMatch(*sSpace*)
  vSpace $\leftarrow$ getValueMappingSpace()
  vMapScore $\leftarrow \infty$ {Stores dissimilarity score for fixed schema match and value mapping}
  while *vSpace* is not empty **do**
    tmpV $\leftarrow$ getNextValueMapping(*vSpace*)
    score $\leftarrow$ computeDissimilarity(*tmpS*, *tmpV*, *T1*, *T2*)
    if *score* $<$ *vmapScore* **then**
      vmapScore $\leftarrow$ *score* {Save current dissimilarity score}
      v1 $\leftarrow$ *tmpV* {Save current value mapping}
    end if
    removeValueMapping(*vSpace*, *tmpV*)
  end while
  if *vmapScore* $<$ *smatchScore* **then**
    s $\leftarrow$ *tmpS* {Store current schema match}
    v $\leftarrow$ *v1* {Store the best value mapping for this schema match}
    smatchScore $\leftarrow$ *vmapScore*
  end if
  removeSchemaMatch(*sSpace*, *tmpS*)
end while

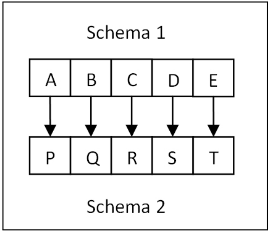
Continuous Schema Matching Algorithm

Algorithm 2 Overview of our approach

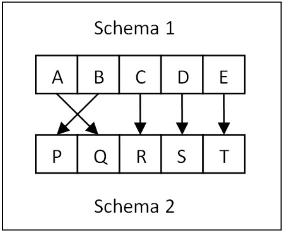
Input: Schemas, *T1* and *T2*
Output: Schema Match *s*, *tmpS*
    sMatchScore $\leftarrow \infty$ { Best Dissimilarity Score}
sSpace $\leftarrow$ getSchemaMatchSearchSpace()
while *sSpace* is not empty **do**
  tmpS $\leftarrow$ getNextSchemaMatch(*sSpace*)
  sMapScore $\leftarrow \infty$ {Stores dissimilarity score for fixed schema match and value mapping for continuous features}
  sMapScore $\leftarrow$ computeDissimilarity(*tmpS*, *T1*, *T2*)
  if *sMapScore* $<$ *sMatchScore* **then**
    s $\leftarrow$ *tmpS* {Store current schema match}
    sMatchScore $\leftarrow$ *sMapScore*
  end if
  removeSchemaMatch(*sSpace*, *tmpS*)
end while

Two-opt Switching

- Simple local search algorithm that we use in our heuristic search strategy
  - Involves swapping of the schema attributes
  - Accepted schema match is one which reduces global objective



An initial schema match



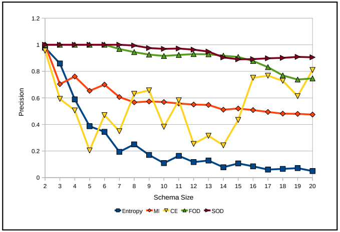
An resulting schema match after 2-opt switching is applied

Pseudocode for heuristic search strategy

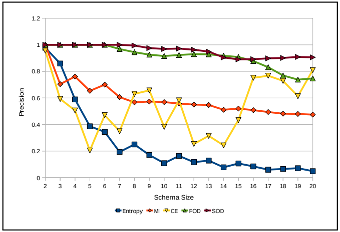
Algorithm 3 Pseudocode for the heuristic search strategy when using two-opt switching.

best_match $\Leftarrow$ get_Initial_Schema_Match()
$D_{Overall} \Leftarrow D^{EU}_{AB}$(best_match)
repeat
  new_match $\Leftarrow$ Two-Opt-Switch(best_match)
  if D^{EU}_{AB}(best_match) $>$ D^{EU}_{AB}(new_match) **then**
    best_match $\Leftarrow$ new_match
    $D_{Overall} \Leftarrow D^{EU}_{AB}$(best_match)
  end if
until no further improvement or a specified number of iterations

Ground-truth Initialization - Categorical

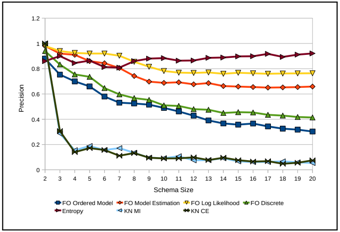


Average precision for CA vs. NY dataset

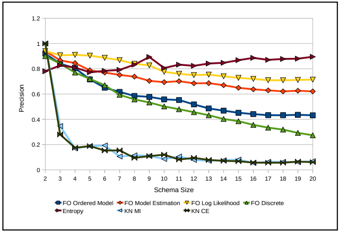


Average precision for CA vs. TX dataset

Ground-truth Initialization - Continuous

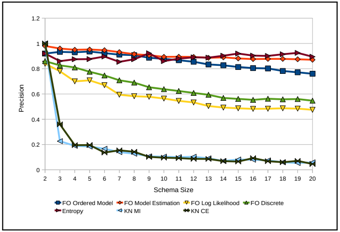


Average precision for CA vs. NY dataset

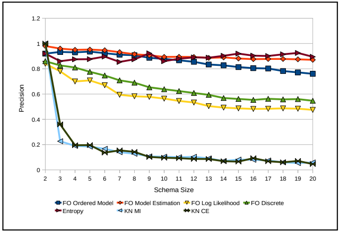


Average precision for CA vs. TX dataset

Ground-truth Initialization - Mixed



Average precision for CA vs. NY dataset



Average precision for CA vs. TX dataset

References I

Fraley, C. and Raftery, A.E.
MCLUST: Software for model-based cluster analysis.
Journal of Classification, 16(2):297–306, 1999.

A. Gal.
Managing uncertainty in schema matching with top-k schema mappings.
Journal on Data Semantics VI, pages 90–114, 2006.

J. Harris and H. Stöcker.
Handbook of mathematics and computational science.
Birkhäuser, 1998.

References II

P. Hoel.
Introduction to mathematical statistics.
Wiley New York, 1962.

A. Jaiswal, D. Miller, and P. Mitra.
Un-Interpreted Schema Matching with Embedded Value Mapping under Opaque Column Names and Data Values.
IEEE Transactions on Knowledge and Data Engineering, 2010.

C. Meek, B. Thiesson, and D. Heckerman.
The learning-curve sampling method applied to model-based clustering.
The Journal of Machine Learning Research, 2:397–418, 2002.

References III

R Development Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2010.
ISBN 3-900051-07-0.