An Iterative Refinement Approach for Mining in Large Sized Data

Dursun Delen,Merylin Kletke, Jinhwa Kim

School of Business
Oklahoma State University

Introduction

- Motivation
- Data mining in Massive Data Set
- Iterative Refinement Method for Massive Data
- Tests and Analysis
- Future Studies

Motivation

Q1.What is a new opportunity in data mining?

- Information overload
- Massive data
- Stream of data

Q2.How to handle this massive data efficiently?

2. Applications of Data Mining

- Stock market prediction
- Bankruptcy prediction
- Feature detection
- Dynamic flexible manufacturing
- Criminal management
- Weather forecasting

Data mining in Massive Data Set

1. What is Massive Data Set?

A data set that is so large and complex that existing methodologies can't cope with it readily.

2. What are the problems with Massive Data Set?

- Size causes aggravation.
  : As the size increases, the accuracy goes down.
- What if data size is keep increasing?

3. Examples of Massive Data Sets

- Prediction with stock market data
- Predicting income with US Census Data
- Earth Observation Systems
- Semiconductor manufacturing
- Crime statistics
- Network traffic
- Health care
- Geographic Information Systems

4. Techniques in Data Mining

- Nonparametric Regression
- Classification
- Clustering
- Neural Networks
- Genetic Algorithms
- CART
- CHAID
- SEE5

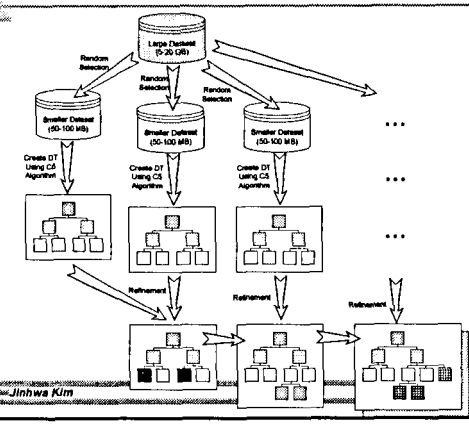
5. Possible Approaches to Massive Data Sets

- Sampling Techniques
- Parallel Neural Networks (Takahash and Hiramatsu 1990, Chiel etc.1992, Owens 2000)
- Merging Knowledge Bases (Gehrke etc. 1998, Hall etc. 1998,Utgoff etc. 1998)
- Distributed Systems (Frieder and Kjell 1991, Paul etc., 1994, Shiroshita etc.1996, George and Knobbe 1998)

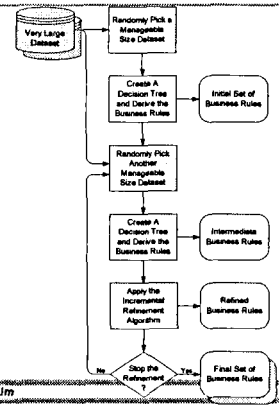
Iterative Refinement Method for Massive Data

- Conceptually it iteratively explores huge data ,collect information and add this information to grow a decision tree.
- Physically it cumulates knowledge from each iteration into a knowledge base.

The Repetitive Methodology of Merging New Rules into the Domain Knowledge Base



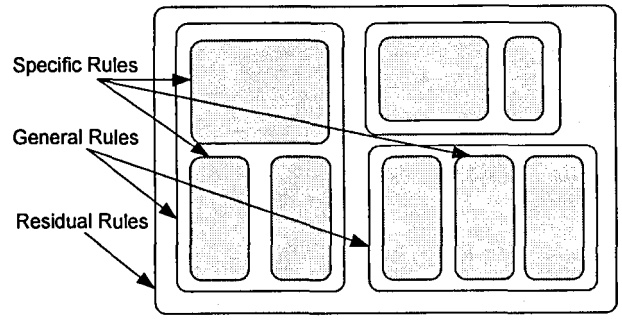
A Visual Display of IRM



Two Steps in IRM

- Step 1:Build the domain knowledge base with a repetitive specialization process.
- Step 2: Refine the domain knowledge base with a repetitive refinement process.

Increasing Coverage Via Specializing Rules



Tests & Analysis

- Data: Census Data 1990
- Independent variables (15)
  age, education, marital status, capital gain, capital loss .. etc.,
- Dependent variable
  income >=$50,000
  income < $50,000

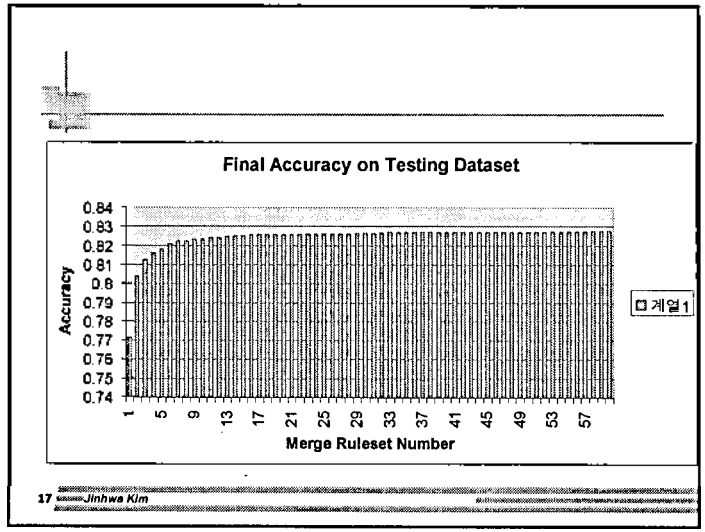
Test results with mid-size (32MB) dataset

- Each training set = 6032 records
- Test set = 15K+ records

Samples	C&RT (w/o Cross Val)			Neural Nets (MLP)			CHAID			Discriminant Analysis		
	<=50K	>50K	Overall	<=50K	>50K	Overall	<=50K	>50K	Overall	<=50K	>50K	Overall
Set #1	89.45%	60.03%	82.22%	77.02%	76.97%	77.01%	98.24%	7.36%	75.91%	93.01%	49.19%	82.24%
Set #2	90.33%	55.03%	81.65%	77.90%	77.89%	77.90%	99.09%	5.33%	76.04%	94.45%	42.65%	81.72%
Set #3	89.51%	55.22%	81.06%	79.27%	78.51%	79.08%	88.41%	48.19%	78.53%	93.40%	50.22%	82.79%
Set #4	89.31%	56.81%	81.33%	79.51%	79.41%	79.48%	92.77%	24.99%	76.14%	93.81%	40.19%	80.64%
Set #5	89.75%	54.51%	81.10%	80.00%	79.08%	79.77%	89.45%	55.35%	81.07%	93.57%	40.84%	80.62%
Mean	89.67%	56.32%	81.48%	78.74%	78.37%	78.65%	93.59%	28.24%	77.54%	93.65%	44.62%	81.60%
St. Dev.	0.40%	2.24%	0.48%	1.24%	0.97%	1.16%	4.91%	22.94%	2.25%	0.53%	4.74%	0.97%
Median	89.51%	55.22%	81.33%	79.27%	78.51%	79.08%	92.77%	24.99%	76.14%	93.57%	42.65%	81.72%
Min	89.31%	54.51%	81.06%	77.02%	76.97%	77.01%	88.41%	5.33%	75.91%	93.01%	40.19%	80.62%
Max	90.33%	60.03%	82.22%	80.00%	79.41%	79.77%	99.09%	55.35%	81.07%	94.45%	50.22%	82.79%

A Test Example - Iterations in IRM

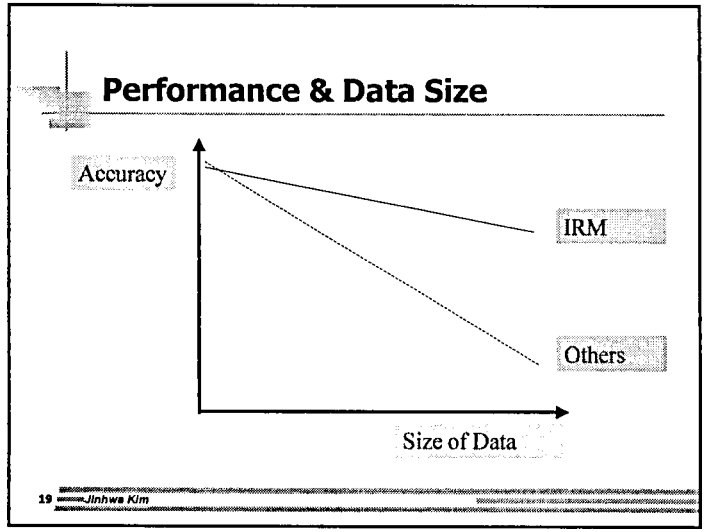
Iterations	Accuracy	Accuracy Improved	Number of Rules	No. New Rules Added
5	82.26	0	167	167
10	84.62	2.36	286	119
15	85.05	0.43	409	123
20	85.16	0.11	511	102
25	85.16	0	637	126
30	85.16	0	637	0
35	85.16	0	637	0
40	85.16	0	637	0
45	85.18	0.02	760	123
50	85.18	0	760	0
55	85.18	0	760	0
60	85.18	0	760	0



Performance Tests

C&RT	Neural Nets	CHAID	Discriminant Analysis	SEE5	SEE5 Seperated & Merged	IRM
81.48	76.65	77.54	81.60	81.82	83.18	85.18

18 Jinhwa Kim



Future Studies

- Huge Data in Congress Library
- Real Time Data Mining
  - On-line stock market data
- Government Applications
  - Transportation
  - Criminal Management

20 Jinhwa Kim