ACSys Cooperative Research Centre for Advanced Computational Systems

Data Mining
The Analysis of Large and Complex Datasets

Graham Williams

CSIRO

THE AUSTRALIAN NATIONAL UNIVERSITY

ACSys Data Mining

- CRC for Advanced Computational Systems
  - ANU, CSIRO, (Digital), Fujitsu, Sun, SGI
  - Five programs: one is Data Mining
  - Aim to work with collaborators to solve real problems and feed research problems to the scientists
  - Brings together expertise in Machine Learning, Statistics, Numerical Algorithms, Databases, Virtual Environments
- Graham Williams, Senior Research Scientist with CSIRO
  Machine Learning
  PhD in Ensemble Decision Tree Induction

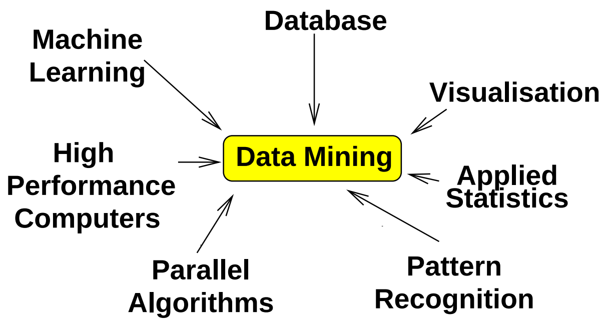
Outline

- The Data Mining Task
- Techniques for Data Mining
- Case Studies in Data Mining
- Research Directions
- Wrap Up

So What is Data Mining?

- *The non-trivial extraction of novel, implicit, and actionable knowledge from large datasets.*
  - Extremely large datasets
  - Discovery of the non-obvious
  - Useful knowledge that can improve processes
  - Can not be done manually
- Technology to enable data exploration, data analysis, and data visualisation of very large databases at a high level of abstraction, without a specific hypothesis in mind.

And Where Has it Come From?



Database

Machine Learning

Visualisation

High Performance Computers

Data Mining

Applied Statistics

Parallel Algorithms

Pattern Recognition

Knowledge Discovery in Databases

- A six or more step process:
  - data warehousing,
  - data selection,
  - data preprocessing,
  - data transformation,
  - data mining,
  - interpretation/evaluation
- Data Mining is sometimes referred to as KDD
- DM and KDD tend to be used as synonyms

Techniques Used in Data Mining

- **Predictive Modelling**
  tree induction, neural nets, regression
- **Database Segmentation**
  clustering, k-means, Kohonen maps,
- **Link Analysis**
  association rules, sequential patterns, time sequences
- **Deviation Detection**
  visualisation, statistics

Typical Applications of Data Mining

- Sales/Marketing
  - Provide better customer service
  - Improve cross-selling opportunities (beer and nappies)
  - Increase direct mail response rates
- Customer Retention
  - Identify patterns of defection
  - Predict likely defections
- Risk Assessment and Fraud
  - Identify inappropriate or unusual behaviour

ACSys Data Mining

Mt Stromlo Observatory

NRMA Insurance Limited

Australian Taxation Office

Health Insurance Commission

Some Research

- Interestingness through Evolutionary Computation
- Virtual Environments
- Data Mining Standards
- Temporal Data Mining
- Spatial Data Mining
- Feature Selection

Outline

- The Data Mining Task
- Techniques for Data Mining
- Case Studies in Data Mining
- Research Directions
- Wrap Up

Outline

- The Data Mining Task
  - History
  - Motivation
  - Disciplines
  - Definitions
  - The Process

Why Data Mining Now?

- Changes in the Business Environment
  - Customers becoming more demanding
  - Markets are saturated
- Drivers
  - Focus on the customer, competition, and data assets
- Enablers
  - Increased data hoarding
  - Cheaper and faster hardware

The Growth in KDD

- Research Community
  - KDD Workshops 1989, 1991, 1993, 1994
  - KDD Conference annually since 1995
  - KDD Journal since 1997
  - ACM SIGKDD http://www.acm.org/sigkdd
- Commercially
  - Research: IBM, Amex, NAB, AT&T, HIC, NRMA
  - Services: ACSys, IBM, MIP, NCR, Magnify
  - Tools: TMC, IBM, ISL, SGI, SAS, Magnify

Outline

- The Data Mining Task
  - History
  - Motivation
  - Disciplines
  - Definitions
  - The Process

The Scientist’s Motivation

- *The Real World*
  - Offers many challenging problems
  - Enormous databases now exist and readily available
- Statistics building models and doing analysis for years?
  - Statistics limited computationally
  - Relevance of statistics if we do not sample
  - There are not enough statisticians to go around!
- Machine Learning to build models?
  - Limited computationally, useful on toy problems, but . . .

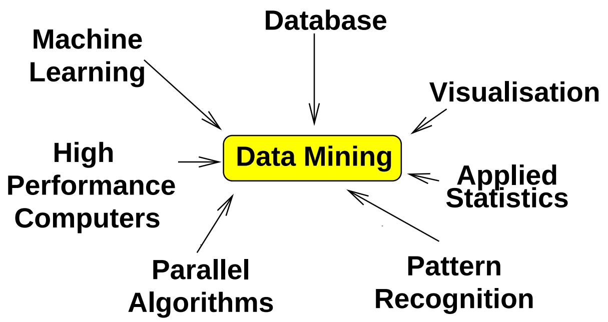
Motivation: The Sizes

- Databases today are huge:
  - More than 1,000,000 entities
  - From 10 to 10,000 entities
  - Giga-bytes and tera-bytes
- Databases a growing at an unprecendented rate
- The corporate world is a cut-throat world
  - Decisions must be made rapidly
  - Decisions must be made with maximum knowledge

Outline

- The Data Mining Task
  - History
  - Motivation
  - Disciplines
  - Definitions
  - The Process

Multiple Disciplines



Databases

- OLTP ⟹ OLAP ⟹ OLAM
- Data Warehouses
  - Subject-oriented, integrated, time-variant, non-volatile
  - Once created then what?
  - Excellent starting point for Data Mining
- Data Marts: Specialised, smaller, data store
- OLAP: Drill-down, roll-up, slice-n-dice, data cubes

From OLAP to Data Mining

- On-Line Analytical Processing
  - Emphasis on Query
  - Generally know what we want to find
  - Expressible in SQL
  - Drill-Down, Data Cubes
- Data Mining
  - Emphasis on Exploration
  - Generally idea of the target but not how to find
  - Let the machine drive the exploration
  - OLAM: Drill-Down, Data Cubes

Statistics

- Data, Counting, Probabilities, Hypothesis testing
- Prediction
  - Friedman: CART, MARS, PRIM
- All important, but no longer by themselves
- Sampling is the antithesis of Data Mining

Machine Learning

- Decision Trees and Classification Rules
- Neural Networks
- K Nearest Neighbours and Clustering
- Genetic Algorithms
- Association Rules
- All useful, but computationaly thirsty

Visualisation

- Exploratory data analysis has long history
- Visualisation in Data Mining
  - Understanding the data
  - Visualising the process
  - Visualising the results of mining

High Performance Computing

- Very large datasets—need rapid access to data
- Very slow algorithms—need to improve
  - Address computational complexity
  - Parallel algorithms

Software Engineering

- SE Practices apply to DM practises
- Developing architectures for Data Mining
- Developing APIs for data and models

Outline

- The Data Mining Task
  - History
  - Motivation
  - Disciplines
  - Definitions
  - The Process

Discovering Knowledge

- *The non-trivial extraction of novel, implicit, and actionable knowledge from large datasets.*
  - Novel = not previously known
  - Implicit = not easily identifiable
  - Actionable = can act upon the discoveries
- Knowledge equates to patterns or models
- Generally we are searching for symbolic descriptions of novel and interesting patterns that are contained in the data. These patterns generally apply to a particular area of the database.

The Key is the Hypothesis

- Data mining "discovers information without a previously formulated hypothesis:" We don't particularly know what we are looking for. We hope to find un-thought of discoveries.
- Compare with:
  - SQL Queries
  - Information Retrieval
  - OLAP
  - Exploratory Data Analysis
  - Visualisation
  - Statistics

Data Mining in a Nutshell

- Start with a raw database
  $\Rightarrow$ RDBMS or Warehouse
- Collect and aggregate variables in a variety of ways
  $\Rightarrow$ dataset $\mathcal{D}$
- Then mine the data: search for patterns implicit in the data and determining whether they are interesting

Formally

- *Dataset* $\mathcal{D}$ of real world *entities* $\mathcal{D} = \{e_1, e_2, \ldots, e_n\}$.
- $\mathcal{D}$ is relational: **R**$(A_1, A_2, \ldots, A_m)$
- Each entity is a tuple $\langle v_1, v_2, \ldots, v_m \rangle$
- The number of attributes m and the number of tuples n are typically "large" (m may be anywhere from 20 to 1000 and n typically greater than 1,000,000)

Generating Candidate Nuggets

- Generate a set of *nuggets* $\mathcal{N} = \{n_1, n_2, \ldots, n_p\}$
- If inducing rules which symbolically describe the nuggets then we equivalently have $\mathcal{R} = \{r_1, r_2, \ldots, r_p\}$
- Rule might consist of conjunction of conditions:
  - $A_i \in [v_1, v_2]$ for numeric attributes
  - $A_i \in \{v_1, v_2, \ldots, v_q\}$ for categorical attributes.

Assessing Interestingness

- Nuggets $\mathcal{N} = \{n_1, n_2, \ldots, n_p\}$
- Generally $p \ll n$ so becoming more amenable to human exploration.
- However, p can still be large (several thousand)
- *How do we measure the interestingness of a nugget?*
- Different approaches employ different measures
- This question we come back to later

Outline

- The Data Mining Task
  - History
  - Motivation
  - Disciplines
  - Definitions
  - The Process

Data Mining is a Process

- A standard six step process is being developed:
  - Business Understanding (25%)
  - Data Understanding (20%)
  - Data Preparaation (25%)
  - Modelling (10%)
  - Evaluation (20%)
  - Deployment

The CRISP-DM

- Being championed by NCR, Daimler-Benz, ISL, OHRA
- **Cr**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining http://www.ncr.dk/CRISP/index.htm
- Define and validate a **Data Mining Process Model**
  - applicable in diverse industry sectors
  - industry and tool neutral
  - large data mining projects executed faster, cheaper, more reliably and more manageably
- Life cycle of six (iterative) phases.

Business Understanding

- Initial phase focuses on understanding project objectives and requirements from a business perspective
- This knowledge is converted into a data mining problem definition
- Develop a preliminary plan designed to achieve the objectives

Data Understanding

- Initial data collection
- Familiarisation with the data
  - identify data quality problems
  - discover first insights into the data
  - detect interesting subsets to form hypotheses for hidden information

Data Preparation

- Construct the mining dataset
- Derived from the initial raw dataset(s)
- Data preparation tasks:
  - table, record, and attribute selection
  - **generation of derived features**
  - data transformation
  - data cleaning

Preparing to Mine

- Data Quality
  - missing data
  - noisy data
  - lead to inconsistent or too general/specific discoveries
- Data Cleaning
  - duplicates
  - inconsistencies
  - identify and merge the same entities

Modelling—The Actual Data Mining

- Select various modelling techniques
- Apply and calibrate modelling techniques
- Typically there are several techniques for the same data mining problem
- Some techniques have specific requirements on the form of data and require stepping back to the data preparation phase

Evaluation

- Evaluate the model and review the steps executed to construct the model
- Does the model properly achieve the business objectives?
- Is there some important business issue that has not been sufficiently considered?
- Decide on the use of the data mining results

Deployment

- Deployment may be:
  - Generate a report of the discoveries made
  - Implement changes in the processes of the organisation
  - Implement a repeatable data mining process
- For successful deployment the customer must understand the actions to be carried out in order to actually make use of the created models

The KDD Process

- An interative process, often requiring multiple loops
- Time consuming process
- The actual mining is one small step in the whole process
- Issues in data management are crucial to success

Outline

- The Data Mining Task
- Techniques for Data Mining
- Case Studies in Data Mining
- Research Directions
- Wrap Up

Data Mining Operations

- **Predictive Modelling** (supervised learning)
  use observations to learn to predict
- **Database Segmentation** (unsupervised learning)
  partition data into similar groups
- **Link Analysis**
  links between individuals rather than characterising whole
- **Deviation Detection**
  outlier detection: becoming a major focus of data mining

Predictive Modelling: Classification

- Goal of classification is to build structures from examples of past decisions that can be used to make decisions for unseen cases.
- Often referred to as supervised learning.
- Decision Tree and Rule induction are popular techniques
- Neural Networks also used

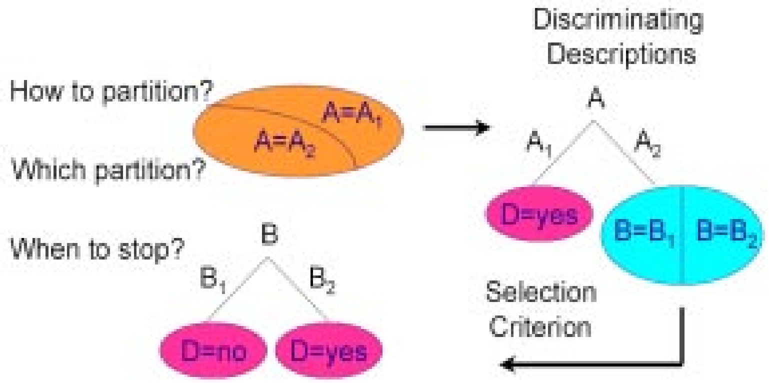
Classification: C5.0

- Quinlan: $\boxed{\textbf{ID3}} \Longrightarrow \boxed{\textbf{C4.5}} \Longrightarrow \boxed{\textbf{C5.0}}$
- Most widely used Machine Learning and Data Mining tool
  Started as Decision Tree Induction, now Rule Induction, also
- Available from http://www.rulequest.com/
- Many publically available alternatives
- CART developed by Breiman et al. (Stanford)
  Salford Systems http://www.salford-systems.com

Decision Tree Induction

- Decision tree induction is an example of a recursive partitioning algorithm
- Basic motivation:
  - A dataset contains a certain amount of information
  - A random dataset has high entropy
  - Work towards reducing the amount of entropy in the data
  - Alternatively, increase the amount of information exhibited by the data

Algorithm



Algorithm

- Construct set of candidate partitions S
- Select best S* in S
- Describe each cell C_i in S*
- Test termination condition on each C_i
  - true: form a leaf node
  - false: recurse with C_i as new training set

Discriminating Descriptions

- Typical algorithm considers a single attribute at one time:
- categorical attributes
  - define a disjoint cell for each possible value: *sex = “male”*
  - can be grouped: *transport* $\in$ *(car, bike)*
- continuous attributes
  - define many possible binary partitions
  - Split $A < 24$ and $A \geq 24$
  - Or split $A < 28$ and $A \geq 28$

Information Measure

- Estimate the gain in information from a particular partitioning of the dataset
- A decision tree produces a message which is the decision
- The information content is $\sum_{j=1}^{m} -p_j \log(p_j)$
  - p_j is the probability of making a particular decision
  - there are m possible decisions
- Same as entropy: $\sum_{j=1}^{m} p_j \log(1/p_j)$.

Information Measure

- $info(T) = \sum_{j=1}^{m} -p_j \log(p_j)$ is the amount of information needed to identify class of an object in T
- Maximised when all p_j are equal
- Minimised (0) when all but one p_j is 0 (the remaining p_j is 1)
- Now partition the data into n cells
- Expected information requirement is then the weighted sum: $info_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i)$

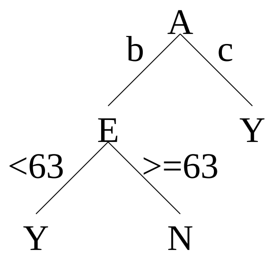
Information Measure

- The information that is gained by partitioning T is then:

$$gain(A) = info(T) - info_x(T)$$

- This *gain criterion* can then be used to select the partition which maximises information gain
- Variations of the Information Gain have been developed to avoid various biases: Gini Index of Diversity

End Result



Pruning

- We may be able to build a decision tree which perfectly reflects the data
- But the tree may not be generally applicable
called **overfitting**
- Pruning is a technique for simplifying and hence generalising a decision tree

Error-Based Pruning

- Replace sub-trees with leaves
- Decision class is the majority
- Pruning based on predicted error rates
  - prune subtrees which result in lower predicted error rate

Pruning

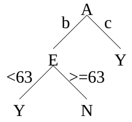
- How to estimate error? Use a separate test set:
  - Error rate on training set (resubstitution error) not useful because pruning will always increase error
  - Two common techniques are cost-complexity pruning and reduced-error pruning
- Cost Complexity Pruning: Predicted error rate modelled as weighted sum of complexity and error on training set—the test cases used to determine weighting
- Reduced Error Pruning: Use test set to assess error rate directly

Issues

- Unknown attribute values
- Run out of attributes to split on
- Brittleness of method—small perturbations of data lead to significant changes in decision trees
- Trees become too large and are no longer particularly understandable (thousands of nodes)
- **Data Mining**: *Accuracy, alone, is not so important*

Classification Rules

- A tree can be converted to a rule set by traversing each path



  - $A = c \Rightarrow Y$
  - $A = b \wedge E < 63 \Rightarrow Y$
  - $A = b \wedge E \geq 63 \Rightarrow N$
  - Rule Pruning: Perhaps $E \geq 63 \Rightarrow N$

Pros and Cons of Decision Tree Induction

- Pros
  - Greedy Search = Fast Execution
  - High dimensionality not a problem
  - Selects important variables
  - Creates symbolic descriptions
- Cons
  - Search space is huge
  - Interaction terms not considered
  - Parallel axis tests only ($A = v$)

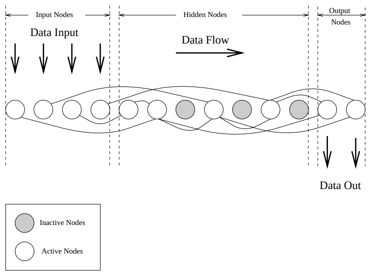
Recent Research

- Bagging
  - Sample with resubstitution from training set
  - Build multiple decision trees from different samples
  - Use a voting method to classify new objects
- Boosting
  - Build multiple trees from all training data
  - Maintain a weight for each instance in the training set that reflects its importance
  - Use a voting method to classify new objects

Predictive Modelling: Regression

- Linear Regression
  - Limited use and biased by outliers
- Non-Linear Regression
  - Difficult with high dimensional data
- Radial Basis Functions
  - More robust and flexible
  - Weighted sum of non-linear functions, each fitted to separate regions
  - Selection of regions is now important

Predictive Modelling: Neural Networks



Nearest Neighbour Approaches

- Compare a new entity with all other entities to find the nearest-neighbours
- Assign the new entity the most popular decision!
- A lazy learner (no generalisation until necessary)

Database Segmentation: Clustering

- Definition:
  - Given n entities find a useful partition into p subsets
- Numerical Clustering
  - Usually easier since distance easy to calculate
  - For example, Euclidean distance
- Conceptual Clustering
  - Targets the derivation of the cluster
  - Symbolic descriptions
  - Can handle non-numeric data

Link Analysis: Association Rules

- A technique developed specifically for data mining
  - Given
    - ∗ A dataset of customer transactions
    - ∗ A transaction is a collection of items
  - Find
    - ∗ Correlations between items as rules
- Examples
  - Supermarket baskets
  - Attached mailing in direct marketing

Determining Interesting Association Rules

- Rules have **confidence** and **support**
  - IF x and y THEN z with confidence c
    - if x and y are in the basket, then so is z in c% of cases
  - IF x and y THEN z with support s
    - the rule holds in s% of all transactions

Example

Transaction	Items
12345	A B C
12346	A C
12347	A D
12348	B E F

- Input Parameters: confidence = 50%; support = 50%
- if A then C: c = 66.6% s = 50%
- if C then A: c = 100% s = 50%

Typical Application

- Hundreds of thousands of different items
- Millions of transactions
- Many gigabytes of data
- It is a large task, but linear algorithms exist

Itemsets are Basis of Algorithm

Transaction	Items
12345	A B C
12346	A C
12347	A D
12348	B E F

Itemset	Support
A	75%
B	50%
C	50%
A, C	50%

- Rule $A \Rightarrow C$
- s = s(A, C) = 50%
- c = s(A, C)/s(A) = 66.6%

Algorithm Outline

- Find all large itemsets
  - sets of items with at least minimum support
  - Apriori and AprioriTid and newer algorithms
- Generate rules from large itemsets
  - For ABCD and AB in large itemset the rule AB ⇒ CD holds if ratio s(ABCD)/s(AB) is large enough
  - This ratio is the confidence of the rule

HIC Example

- Associations on episode database for pathology services
  - 6.8 million records X 120 attributes (3.5GB)
  - 15 months preprocessing then 2 weeks data mining
- Goal: find associations between tests
  - cmin = 50% and smin = 1%, 0.5%, 0.25%
    (1% of 6.8 million = 68,000)
  - Unexpected/unnecessary combination of services
  - Refuse cover saves $550,000 per year

Outline

- The Data Mining Task
- Techniques for Data Mining
- Case Studies in Data Mining
- Research Directions
- Wrap Up

The Hot Spots Methodology

- Search for interesting areas within a very large database
  - How to partition the data?
    Finding the nuggets
  - How to assess interestingness?
    Weighing the nugget
- In our case studies we have found clustering and rule induction to work together well to identify nuggets
- We are still novices in properly weighing the nuggets

Mining the Knowledge Mine

- Data-driven approach to generate the rules
  - Cluster $\mathcal{D}$ into a set of subsets
  - Describe each subset using Classification
  - To generate nuggets $\mathcal{R} = \{r_1, r_2, \ldots, r_p\}$.
- While p is generally much smaller than n ($p \ll n$) it can still be substantial (perhaps one or two thousand for n in the millions). p generally remains too large for manual consideration.
- *Manual and time-consuming process of analysing lots of statistics about the groups that “looked” interesting.*

ACSys Data Mining

Mt Stromlo Observatory

NRMA Insurance Limited

Australian Taxation Office

Health Insurance Commission

Macho

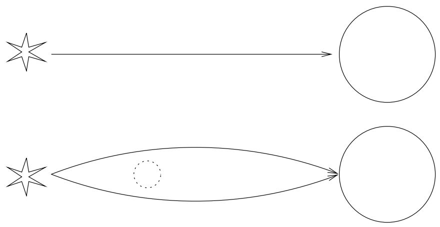
MAssive Compact Halo Objects

The search for the dark matter of the Universe

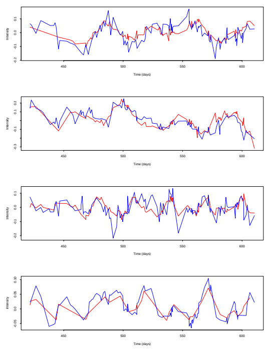
20 million stars recorded nightly for over 4 years

Search for micro lensing events

Micro-Lensing



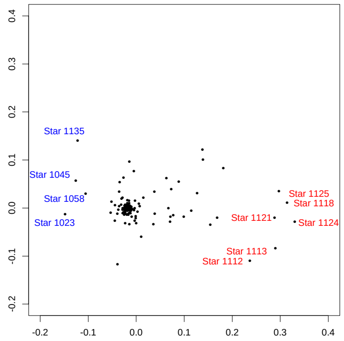
Sample Star Light Curves



Transformed into Feature Space

- Time Series mapped into Feature Space
  Fourier Transforms, Auto-Regression, Wavelets, . . .
- Cluster stars using feature space
- Investigate outliers

Approach: Cluster then Describe then Measure

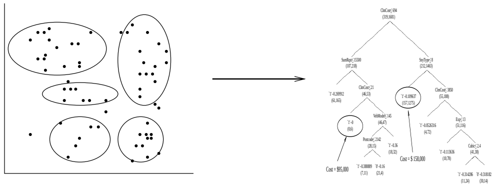


NRMA: Motor Vehicle Insurance

Insurance premium setting and risk rating

- Actuaries study data and domain to define general formula
- Several million transactions annually
- Better understand dynamics of the areas of risk
- Consider more than the traditional small number of factors

Cluster then Describe then Measure



Find the Interesting Groups

Rule 1 NCB $<$ 60 **and** Age $\leq$ 24 **and** Address is Urban
Rule 23 Age $>$ 57 **and** Vehicle $\in$ {Utility, Station Wagon}

Nugget	*Claims*	*Total*	*Proportion*	*Average Cost*	*Total Cost*
1	150	1400	11	3700	545,000
2	140	2300	6	3800	535,000
3	5	25	20	4400	13,000
4	10	120	8	7900	79,100
5	20	340	6	5300	116,000
6	65	520	13	4400	280,700
7	5	5	100	6800	20,300
. . .					
60	800	1400	5.9	3500	2,800,000
All	3800	72000	5.0	3000	12,000,000

Finding the Interesting Groups

Evaluate the **large** collection of groups (or Hot Spots) to find those that are important to the core business.

Nugget	*By Claims*	*By Proportion*	*By Average Cost*
2	Y		
3			Y
19			Y
24		Y	
34	Y	Y	Y
35	Y		Y
36			Y
40	Y		Y

Find the Interesting Groups

Rule 1 NCB $<$ 60 **and** Age $\leq$ 24 **and** Address is Urban
Rule 23 Age $>$ 57 **and** Vehicle $\in$ {Utility, Station Wagon}

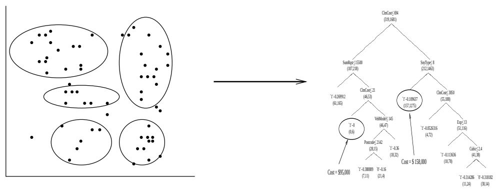
Nugget	*Claims*	*Total*	*Proportion*	*Average Cost*	*Total Cost*
1	150	1400	**11**	**3700**	545,000
2	140	2300	6	**3800**	535,000
3	5	25	**20**	**4400**	13,000
4	10	120	8	**7900**	79,100
5	20	340	6	**5300**	116,000
6	65	520	**13**	**4400**	280,700
7	5	5	100	**6800**	20,300
. . .					
60	800	1400	5.9	**3500**	2,800,000
All	3800	72000	5.0	3000	12,000,000

Health Insurance Commission

Medicare

- Terabytes of patient claims since the inception of Medicare
- Inappropriate Provider practices an ongoing focus
- Exploration of *public fraud* (including doctor shoppers)
- Exploration of the practise of pathology

Cluster then Describe then Measure

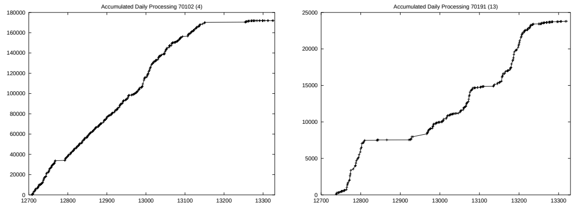


Cluster then Describe then Deliver

Rule 1	Age is between 18 and 25 **and** Weeks $\geq$ 10
Rule 2	Weeks $<$ 10 **and** Benefits $>$ \$350

ugget	Size	Age	Gender	Services	Benefits	Weeks	Hoard	Regular
1	9000	30	F	10	30	2	1	1
2	150	30	F	**24**	**841**	4	2	4
3	1200	65	M	7	220	20	1	1
4	80	45	F	**30**	**750**	10	1	1
5	90	10	M	12	**1125**	10	**5**	2
6	800	55	M	8	550	7	1	**9**
. . .								
280	30	25	F	15	450	15	2	6
All	40,000	45		8	30	3	1	1

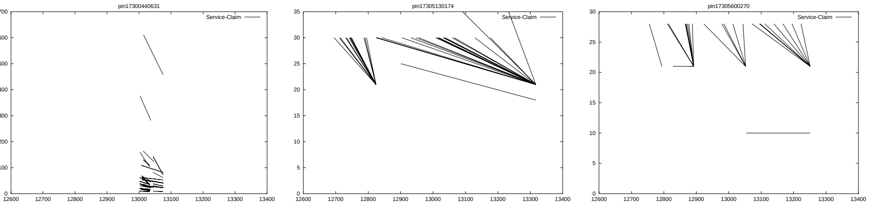
HIC Time Plots



But there may be several thousand of these.

Claim Hoarders

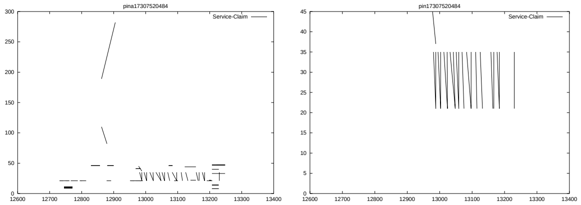
A distinct group of behaviour identified as Claim Hoarders



But there may be many millions of these.

Medicare Regulars

Group of patients with very regular activity:



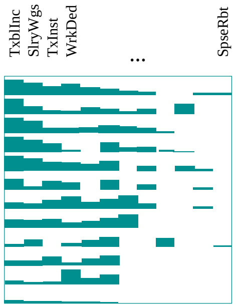
Remove non-cash payments!!!

Australian Tax Office

- Fraud Prevention and Control
- Discover nuances that can establish tax payer has non-legitimate (possibly fraudulent) claims
- Conventional statistical techniques not working: more than looking for averages, outliers, etc.
- Effective selection of possible hits (not too many)

ATO

Identifying a tax avoidance scheme



Hot Spot	Size
1	238
2	1221
3	25 534
4	55
5	19
6	1184
7	198 234
8	4356
9	1417
10	14
11	8
12	473

Outline

- The Data Mining Task
- Techniques for Data Mining
- Case Studies in Data Mining
- Research Directions
- Wrap Up

Interestingness

- We are drowning in a wealth of “discoveries”
- Visualisation is not enough—VR beginning to help.
- Still need a better way to help us *interactively* explore through the space of possible groupings of the data.
- Combined with VR to provide a data driven yet user controlled exploratory data mining system
- We need a way of better exploring the space and identifying interestingness more formally.

Fitness = Interestingness

- We cast our problem to that of searching for a collection of data subsets (described by rules) which we find interesting
- Our rule set $\mathcal{R}$ can serve as a starting point for the search (although randomly generated starting points could be used).
- If we can define what we mean by interesting then we could use that as out fitness measure to evolve a fitter collection of rules.

Rule Induction

- Rule Evolution is not new, but
  We do not seek rule accuracy in the usual sense
  We do not seek coverage in the usual sense
- Instead we seek a collection of rules that describe different regions of the database and are interesting.

Specifying Interestingness

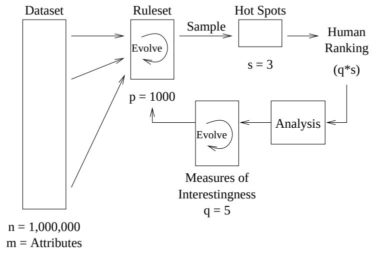
- Specifying interestingness is now the difficult task.
- There are many aspects to take into account.
- The domain user will essentially refine their idea of interestingness as they proceed in their exploration
- Keep the domain user in the look: In an evolutionary sense, maintain them as the oracle to decide fitness of the interestingness measure

How to Capture Interestingness

- We can capture interestingness as some formula involving:
  - Distance
    A measure of how far a group prototype is from the population prototype.
  - Template
    A pattern that is used to increase the interestingness measure.
  - Belief
    Well know truths of the domain to check for surprising discoveries.
  - Sizes
    Size of corresponding dataset and complexity of rule

An Architecture for Evolutionary Hot Spots

n a formal language for expressing interestingness we can now explore the evolution of fitness measure itself, allowing the data driven discoveries to be assessed by the domain r, whose input is captured to evolve measures of interestingness which are then employed volve new nuggets!



Other Research Directions

- Virtual Environments
- Data Mining Standards
- Temporal Data Mining
- Spatial Data Mining
- Feature Selection

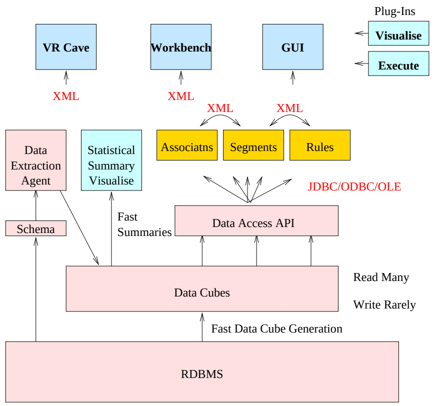
Virtual Environments

- The screen is too limited in what it can present visually
- Virtual environments provide considerably more real-estate
- Explore the results of data mining
- Rely on the significant human ability for perception
  - Pick up patterns in the forrest of trees that catch the eye

Standardisation Issues

- Standard Architectures
- Standard Interfaces
  - Data Access Interfaces
  - Model Communication Interfaces

Standard Architecture



Plug-Ins
Visualise
Execute
VR Cave
Workbench
GUI
XML
XML
XML
XML
Data Extraction Agent
Statistical Summary Visualise
Associatns
Segments
Rules
JDBC/ODBC/OLE
Schema
Fast Summaries
Data Access API
Data Cubes
Read Many
Write Rarely
Fast Data Cube Generation
RDBMS

Temporal Data Mining

- Time is an issue in many databases
- Most OLTP systems are time-based transactional systems
- Yet, we tend to factor time out to do Data Mining
- How can we use time directly in Data Mining?
  - Temporal Logics
  - Temporal Databases

Spatial Data Mining

- Spatial Inforation Systems (and GIS) now common technology
- Spatial Databases store Spatially Indexed data
- Again, in Data Mining we tend to factor space out!
- How can we use spatial information directly in Data Mining?

Feature Selection

- *Once we have the right features data mining is trivial!*
- But how do we best aggregate the source features?
  - Aggregate transactions in some way:
    number of transactions every 3 months
    number of different types of transactions
  - Aggregate spatial data appropriately:
    in particular census collection disctrict
- But the aggregations encapsulate preconceived ideas

Outline

- The Data Mining Task
- Techniques for Data Mining
- Case Studies in Data Mining
- Research Directions
- Wrap Up

Outline

- Wrap Up
  - Tools
  - Privacy

The Data Mining Vendor Space

- Visit http://www.cmis.CSIRO.AU/Graham.Williams/dataminer/Catalogue.html
- Over 60 vendors in the market today
  - A crowded market
  - Tool Vendors versus Service Providers
  - Sales of less than 30!
- Market Share: IBM, ISL, SAS, SGI, TMC

IBM: Intelligent Miner

- A collection of tools to handle organisational/transactional data for Data Mining.
- Some of the best Data Mining research from IBM research labs, but not all of it captured in Intelligent Miner
- **Techniques**: Association, Segmentation, Time Sequence, NN, Rules
- Client/Server with Java-based GUI—reasonably intuitive

ISL: Clementine

- Was first with an excellent point-n-click interface for Data Mining
- Took the drudgery out of linking lots of tools together
- Under the bonnet, standard collection of techniques, including C5.0, NN, clustering, associations, etc.
- Newest release addressing standardisation issues

SAS: Enterprise Miner

- Integration with the SAS world is a big benefit
- Standard techniques: Clustering, Decision Trees, Linear and Logistic Regression, Neural Networks
- Still in beta and seeking stability

SGI: MineSet

- Hot on visualisations
- Excellent for exploring the data to understand it
- Contains usual collection of basic tools
- Allows other tools to plug-in (e.g., AutoClass from UltiMode)

TMC: Darwin

- The “first” data mining tool to market?
- Early Versions
  - Connection Machine
  - Provided NN and Decision Trees
- Current Versions
  - Client (NT) / Server (Unix)
  - Ease-of-use
  - More model builders

Privacy and Data Mining

- Laws in many countries directly affect Data Mining, and it is worth being aware of them—penalties are often severe.
- The OECD Principles of Data Collection are relevant.

OECD Principles of Data Collection

- **Collection limitation**:
  Data should be obtained lawfully and fairly, while some very sensitive data should not be held at all
- **Data quality**:
  Data should be relevant to the stated purposes, accurate, complete and up-to-date; proper precautions should be taken to ensure this accuracy
- **Purpose specification**:
  The purposes for which data will be used should be identified, and the data should be destroyed if it no longer serves the given purpose

OECD Principles of Data Collection

- **Use limitation**:
  Use of data for purposes other than specified is forbidden, except with the consent of the data subject or by authority of law
- **Security safeguards**:
  Agencies should establish procedures to guard against loss, corruption, destruction, or misuse of data
- **Openness**:
  It must be possible to acquire information about the collection, storage, and use of personal data

OECD Principles of Data Collection

- **Individual participation**:
  The data subject has a right to access and challenge the data related to him or her
- **Accountability**:
  A data controller should be accountable for complying with measures giving effect to all these principles

Further Information

- *http://www.cmis.csiro.au/Graham.Williams/DataMiner*
- **Discovering Data Mining** by Cabena, Hadjinian, Stadler, Verhees, and Zanasi. Prentice Hall, 1998.
- **Data Warehousing, Data Mining, and OLAP** by Berson and Smith. McGraw-Hill, 1997
- **Machine Learning** by Mitchell. McGraw-Hill, 1997