Week 1 Lecture 2

Unit Coordinator - Dr Liwan Liyanage

School of Computing, Engineering and Mathematics

Discovering Knowledge in Data

Introduction to Data Science

Following slides introduce the process of data mining and its six phases. Five case studies are provided to explain the six phases and how it relates to each other within the CRISP-DM process for data mining.

WESTERN SYDNEY
UNIVERSITY

CRISP - DM process

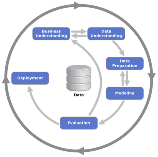

Figure 1:

From *datasciencecentral.com.*

Introduction to Data Mining

Examples of Data Mining

- Bank of America
  - 13 million contact Bank of America's call center each month
  - In past, customers listened to same marketing message
  - Whether relevant to customer or not
  - Kelly, VP Database Marketing states, "*. . . we want to be as relevant as possible to each customer*"
  - Customer profiles available to service representatives
  - May suggest applicable products or services
  - Data mining helps identify marketing approach based on customer's profile

Introduction to Data Mining ctd...

- Homeland Security
  - Shortly after 9/11/2001 events, FBI announced identification of five terrorists in consumer database records
  - One had 30 credit cards with $250,000 debt
  - Another had 12 different addresses
  - Former President Clinton concluded data should be proactively searched
  - Clinton said, "*...they have 12 homes, they're either really rich or up to no good...shouldn't be hard to figure out which.*"

Introduction to Data Mining ctd…

- Gene Expression Database
  - In children, brain tumors represent deadly form of cancer
  - 3,000 cases diagnosed per year
  - Children's Memorial Hospital building gene expression database
  - Goal is developing more effective treatment
  - Bremer, Director of Brain Research, uses Clementine as initial step in tumor identification
  - Classification identifies one of 12 different tumor types

Case Study 1

Analyzing Automobile Warranty Claims

- Business Understanding
    - Objectives include improving customer satisfaction and reducing costs
    - Manufacturing engineers consulted to formulate business problems
    - Data mining techniques used to uncover possible issues:
        - Warranty claim interdependencies?
        - Past claims associated with future claims?
        - Association between claim and repair facility?

Case Study 1 ctd. . .

- Data Understanding
  - 40GB QUIS database containing 7 million vehicle records used
  - Vehicle records include manufacturing location, warrant claims, and additional codes
  - Database unintelligible to non-domain experts
  - Costly effort to consult with domain experts from different departments
- Data Preparation
  - QUIS discovered to have limited SQL access
  - Cases and variables manually extracted
  - Additional variables derived for modeling phase
  - Proprietary data mining software used
  - Data format requirements varied for different algorithms
  - Resulted in exhaustive pre-processing of data

Case Study 1 ctd. . .

- Modeling Phase
  - Applied Bayesian networks and association rules to uncover dependencies between warranty claims
  - Discovered specific combination of construction specifications doubles probability of electrical cable claim
  - Investigated whether some garages had more claims than others
  - Remaining results confidential

Case Study 1 ctd. . .

- Evaluation
  - Researchers disappointed in results
  - Association rules could not be generalized
  - Rules “not interesting” according to domain experts
  - Data models fell short of business objectives
  - Legacy databases not suited to data mining
  - Proposal suggested database redesign for future data mining efforts
- Deployment
  - Foregoing effort identified as pilot project, models not deployed
  - Future data mining efforts planned to integrate more closely to database systems at DaimlerChrysler

WESTERN SYDNEY
UNIVERSITY

Case Study 1 ctd…

- Summary
  - Uncovering hidden nuggets very difficult
  - During each phase, researchers encountered roadblocks
  - Applying new data mining effort problematic
  - Data mining effort requires management support
  - Substantial human participation required at every stage
  - Installation, configuration, and data mining modeling not magic
  - Wrong analysis leads to possibly expensive policy recommendations
  - No guarantee data mining effort delivers actionable results
  - However, used properly, data mining may provide profitable results

Case Study 2

Predicting Abnormal Stock Market Returns

- Business Understanding
  - Alan Safer reports stock market trades by insiders have abnormal returns
  - Profits by outsiders can be increased, by using legal insider trading information
  - Safer attempts to predict abnormal stock price returns
- Data Preparation
  - Rank of company insiders not considered important
  - Also, company insiders omitted where not involved in company decisions

Case Study 2 ctd...

- Modeling
  - Data split training = 80% and validation = 20%
  - Neural Network model uncovered results:
  - (a)Several groups had most predictable abnormal returns:
    - Electronic equipment, excluding computer equipment
    - Chemical products
    - Transportation equipment
    - Business services
  - (b)Predictions looking farther into future increased ability to predict unusual variations
  - (c)Abnormal stock returns of small companies easier to predict

Case Study 2 ctd...

- Evaluation
  - Multivariate Adaptive Regression Spline (MARS) also applied to data
  - Uncovered similar findings to Neural Network model
  - Confluence of results is powerful method of evaluating validity of model
  - This increases confidence in results
- Deployment
  - Safer published findings in Intelligent Data Analysis

Case Study 3

Mining Association Rules from Legal Databases

- Business Understanding
  - Ivkovic, Yearwood, and Stranieri mine association rules from large database of applicants for government-funded legal aide in Australia
  - Legal data highly unstructured
  - Goal is to improve delivery of legal services
- Data Understanding
  - Data provided by Victoria Legal Aid (VLA)
  - Contains 380,000 applications with 300 attributes
  - Domain experts consulted in effort to reduce dimensionality
  - Researchers selected seven of the most important inputs for inclusion in data set
- Data Preparation
  - VLA data set relatively clean
  - VLA database administration system responsible for high-quality data

Case Study 3 ctd. . .

- Modeling
    - Rules restricted to having both single antecedent and consequent
    - Many interesting and uninteresting rules uncovered
    - Researchers adopted premise that interesting rules spawn interesting hypotheses
    - For example, possible reasons for rule “If place of birth = Vietnam, then law type = criminal law” include:
        - Vietnamese applicants applied for criminal law assistance only
        - Vietnamese applicants committed more crimes than other groups
        - Perhaps high proportion of males applied, and males more closely associated with criminal activity
        - Vietnamese didn’t have access to VLA promotional material
    - Researchers concluded first hypothesis most likely
    - Note intense human activity required for data mining process

WESTERN SYDNEY UNIVERSITY

Case Study 3 ctd...

- Evaluation
  - Three domain experts estimated confidence level of 144 rules
  - These estimates compared to confidence level reported by software
- Deployment
  - Web-based application developed (WebAssociator)
  - Non-specialists able to access rule-building engine
  - Researchers suggest WebAssociator deployment to enhance judicial system
  - May identify unjust processes

Case Study 4

Predicting Corporate Bankruptcies using Decision Trees

- Business Understanding
  - Recent economic crisis has spawned many corporate bankruptcies in East Asia
  - Sung, Chang, and Lee developing models predicting bankruptcies that maximize interpretability of results
  - Therefore, decision trees used for modeling
- Data Understanding
  - Data included two groups of Korean companies
  - Those that went bankrupt 1991-1995 during stable period
  - Companies that went bankrupt during crisis years 1997-1998
  - 29 firms identified, mostly manufacturing
  - Financial data collected from Korean Stock Exchange

WESTERN SYDNEY
UNIVERSITY

Case Study 4 ctd. . .

- Data Preparation
  - 56 financial ratios identified through literature search
  - 16 dropped due to duplication
  - Measures included growth, profitability, etc.
- Modeling
  - Separate decision tree models were applied under “normal” and “crisis” conditions
  - Normal-conditions rules uncovered:
    - If productivity of capital > 19.65, then predict non-bankrupt with 86% confidence
    - If productivity of capital <= 19.65 and ratio of cash flow to total assets <= -5.65, then predict bankrupt with 84% confidence
  - Crisis-conditions rules uncovered:
    - If productivity of capital > 20.61, predict non-bankrupt with 95% confidence
    - If ratio of cash flow to liabilities > 2.54, predict non-bankrupt with 85% confidence
  - “Cash flow” and “productivity of capital” important predictors, regardless of economic conditions

Case Study 4 ctd. . .

- Evaluation
  - Panel of domain experts concluded “productivity of capital” most important attribute for differentiating firms at risk
  - Domain experts verified results of decision tree
  - Group confirmed results would generalize to population of Korean manufacturing firms
  - Discriminant analysis determined many of the 40 financial ratios were important predictors
- Deployment
  - No specific deployment took place
  - However, financial institutions in Korea became more aware of important predictors of bankruptcy

Case Study 5

Profiling the Tourism Market using k-Means Clustering Analysis

- Business Understanding
  - Hudson and Richie were interested in studying intra-province tourism behavior in Alberta, Canada
  - Goal was development of marketing campaign for tourism in Alberta (sponsored by Travel Alberta)
  - Models created in effort to quantify factors for choosing vacations in Alberta
- Data Understanding
  - Data collected from 13,445 Albertans using phone survey in 1999
  - Only 3,071/13,445 records included in modeling

Case Study 5 ctd. . .

- Data Preparation
  - One question asked respondents to indicate which of 13 factors most influenced their travel decisions
  - Factors included accommodations, weather conditions, etc.
- Modeling
  - Between two and six clusters explored with k-Means
  - Five-cluster solution chosen with profile names:
    - Young Urban Outdoor Market
    - Indoor Leisure Traveler Market
    - Children-first Market
    - Fair-weather-friends Market
    - Older, Cost-conscious Traveler Market

Case Study 5 ctd…

- Evaluation
  - Discriminant analysis verified “reality” of clusters
  - Classified 93% correctly
- Deployment
  - Findings resulted in launch of new campaign, “Alberta, Made to Order”
  - More than 80 projects launched
  - Travel Alberta found increase of 20% in number of Albertans considering Alberta “top-of-the-mind” travel destination