KDD Workshop, 2005
Chicago, Illinois August 22-25, 2005

Generation of Synthetic Data Sets for Evaluating the Accuracy of Knowledge Discovery Systems

Daniel Jeske[1,3], Behrokh Samadi[2], Pengyue J. Lin[4], Lan Ye[3], Sean Cox[4], Rui Xiao[1], Ted Younglove[3], Minh Ly[3], Doug Holt[4], and Ryan Rich[4]

[1]*Department of Statistics, College of Natural and Agricultural Sciences, University of California, Riverside, CA 92521*
[2] *Lucent Technologies, Performance Analysis Department, Holmdel, New Jersey, 07733*
[3] *UCR Statistical Consulting Collaboratory, Riverside, CA, 92521*
[4] *College of Humanities, Arts, and Social Sciences, University of California, Riverside, CA 92521*

Introduction

Information Discovery and Analysis Systems (IDAS) designed to use to identify potential events that could occur in the future.

Need to have data sets available to guide the development and implementation of data mining techniques and to test the accuracy of the IDAS.

Developing test cases for an IDAS requires background data sets into which hypothetical future scenarios can be overlaid. The IDAS can then be measured in terms of false positive and false negative error rates.

Obtaining real test data sets can be an obstacle due to both privacy issues and also the time and cost associated with collecting multiple instances of a diverse set of data sources.

In this poster, we give an overview of the design and architecture of an IDAS Data Set Generator (IDSG) that will enable a fast and comprehensive test of an IDAS.

A credit card transaction example illustrates the use of semantic graphs to capture attribute dependencies, and also illustrate our statistical approach for approximating high dimensional multivariate distributions.

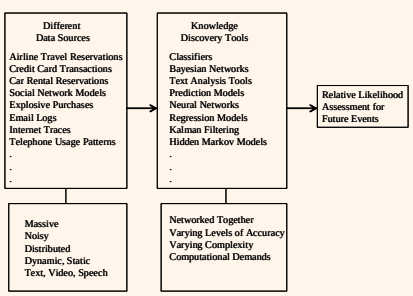
IDSG Goals

IDSG is a synthetic data generation tool designed to create data for testing IDAS.

An IDAS utilizes various and diverse data sources manipulated with knowledge discovery tools to piece together evidence about a future event.

Nature of the evidence depends on the type of data sources utilized by the data mining tools.

Validity of the projected event depends on the quality of the data sources and the type of data mining techniques utilized.



Information Discovery Analysis System

Challenges in Generating Test Case Data

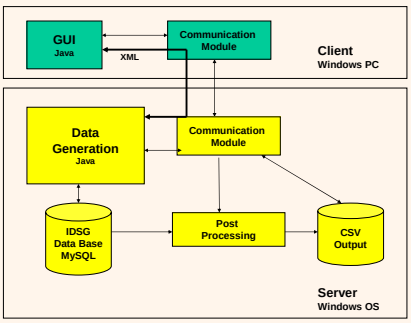
1) Proprietary issues and/or privacy rights
2) Practical challenges (time and costs) of amassing diverse data from multiple sources
3) Limited availability of 'real data' to train predictive models for high dimensional data.
4) Multiple types of data with varying formats and sizes

Guiding Philosophy

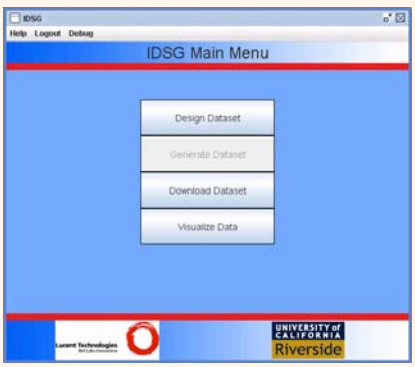
Our goal is not to attempt generation of synthetic data that is virtually indistinguishable from real data, but rather to generate data sets that are *sufficient for evaluation purposes*.

IDSG Tool

User Performs the Following Functions:
- selects from a list of developed data set applications
- enters appropriate data generation parameters and/or filters
- specifies output file formats



IDSG Client Server Architecture



IDSG Main Menu

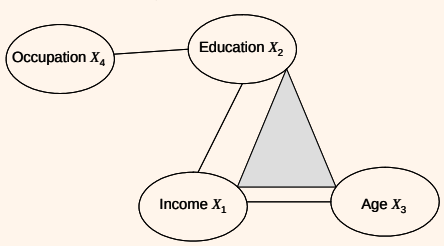
Semantic Graph Representations

Semantic graph shows the availability of information about variable associations.

The graph below shows that two-way information (e.g., bivariate distributions) is available for (Occupation, Education), (Education, Income) and (Income, Age)

In addition, 3-way information (e.g., trivariate distribution) is available for (Education, Income, Age)

Higher dimensional joint distributions are unlikely to be available from low-cost and easily obtained data sources.



Semantic Graph Depiction of Available Information on Associations

Credit Card Transaction Application

People Attributes

Name, Address, Telephone, Email, Gender, Age, Ethnicity, US Citizenship, Marital Status, Social Security Number, Drivers License Number, Vehicle Information (Make, Model, Year, License Plate Number), Occupation, Education

Credit Card Transaction Attributes

Credit Card Numbers, Credit Card Activity Window, Credit Card Purchase Amounts, Credit Card Expense Categories

Data Generation Algorithms

Attributes are generated using one of the following alternative schemes:

1) Resampling

Names, Address, Social Security Number

For example, ethnicity and name are jointly sampled from a public list. SSN is sampled for the Social Security Death Master Index

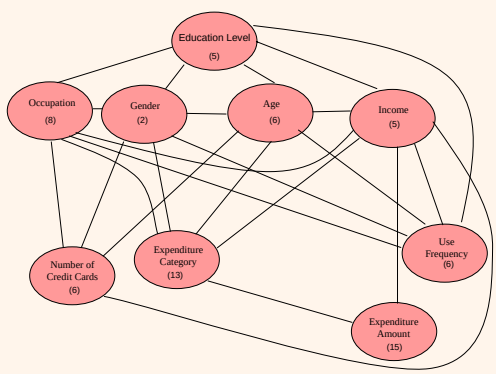
2) Rule-Based

Email address, Drivers License Number, License Plate Number, Vehicle Information Number, Telephone Number, Credit Card Numbers

For example, Drivers License Numbers follow specific state formats, Credit Card Numbers are generated with specific rules for AMEX, Visa, etc.

3) Semantic Graph and Iterative Proportional Fitting Algorithm

Education Level, Occupation, Gender, Age, Income, Number of Credit Cards, Expenditure Category, Credit Card Use Frequency and Expenditure Amount are all derived by applying IPF to the Semantic Graph below:



Semantic Graph for Credit Card Application

Each line between a pair of attributes indicates the availability of a two-way marginal distribution.

The IPF algorithm fits the joint distribution for all 9 variables by starting with a uniform joint distribution.

In an iterative fashion, the 9-dimensional joint distribution is modified so that after the iterations converge, the implied two-way marginals for all pairs of variables shown above agree with the available inputted information.

Research Funded By

Department of Homeland Security Grant HO32040450