CRM Data Mining: Methods of Dimensionality Reduction and Choosing A Right Technique

Dr. Nethra Sambamoorthi, Ph.D
CRMportals Inc.,
11 Bartram Road
Englishtown, NJ 07726

[Key words: variable reduction, space partitioning, regression methods, clustering methods, neural networks, genetic algorithm, tree based methods, decision rules, logistic regression]

The data mining methods are methods to mine voluminous data to find gold nuggets (decisions) from data. The decisions could be as simple as to making decision for a particular customer, to decision as to what assets of a losing company could be leveraged to build the competitive case.

The dimensionality reduction comes down to basically reducing the number of variables to few variables, category (newly definable) variables (reduce huge dimensions to linear or non-linear combinations of variables, also called reduce variables to category dimensions) or categorizations of huge dimensional spaces into understandable fewer partitioned spaces, category spaces (reduce huge dimensions to category spaces), with appropriate discounting of unusual dimensions, variables, categories, and spaces (also termed outlying variables, outlying categories, and outlying spaces) and trends not attributable to robust decision rules from the data.

In a large credit card company the number of variables to analyze are thousands. The ultimate question is whether the individual could be authorized for upgrade of his card to higher priced, better-serviced card with in the franchise.

The analysts could use all the following: However based on the purpose, whether we are looking for decision rules out of the data, how easy to interpret, or how robust (less affected by outlying observations and probability distributional assumptions of data), one may choose the right one. Some guidelines are provided below to that effect.

Type	Variable reduction or space partitioning	Decision method or exploratory	How fast the method is?	Is it less or more robust?	Is importance of a variable known?
Discriminant analysis	Variable reduction and space partitioning	Decision	Fast with right variable search method	Less	Known
Logistic regression	Variable reduction and space partitioning	Decision	Fast with right variable search method	More	Known
Principle components	Variable reduction	Exploratory/	Fast	Less (non-linearity not addressed)	Known
Factor analysis	Variable reduction	Exploratory	Fast	Less (non-linearity not addressed)	Known
Cluster analysis	Space partitioning	Decision	Fast with right variable search method; many variations	More	No
Exploratory projection pursuit	Variable reduction	Exploratory	Fast	Less (non-linearity not	Knowable but of no use

				addressed)	
Tree based methods	Space partitioning	Decision	Fast	More	Yes
Neural networks	Variable and Space partitioning	Decision	Fast; prep time and interpretation is time consuming; too much of specific domain knowledge about neural networks is required	More	No
Nearest neighbor method	Space reduction	Decision	Fast	More	No
Genetic algorithm	Variable reduction and space partitioning	Decision	Fast – prep time and interpretation is time consuming	local optima are likely obtained	Knowable but of little use

Based on this, it is clear why certain methods (methods highlighted with blue color) are more popular than others. Most of the problems are solvable with these three methods, though there are some specific situations that require other special methods. For example, in an engineering problem with complex data types, it is lot more easier to use neural networks than other types.