Data mining techniques for actuaries: an overview

Emiliano A. Valdez

joint work with Banghee So and Guojun Gan

University of Connecticut

Advances in Predictive Analytics (APA) Conference
University of Waterloo, Canada

1-2 December 2017

Data mining

- Refers to a computational process of exploring and analyzing enormous amount of data to uncover hidden and useful information.
- Information for what: to process and efficiently reduce data into a more summarized, analytical and interpretable representation.
- The complicated process for data mining: data acquisition, data preparation, preliminary exploration, data modeling and model evaluation.
- What is it generally use for: to deliver predictive models applicable to new data.
- Predictive models are becoming increasingly important for actuaries in all areas of insurance and financial security programs: life, health, pensions, property and casualty.

Our goals and advantages of data mining

- Not surprisingly, data mining has been applied in several industries: social media, manufacturing, retail, banks and lenders, and government.
- Our goals are to give an overview of data mining techniques and how some of these techniques can be applied to some actuarial problems.
- Previous work: Guo (2003)
- Some advantages of data mining techniques (Fürnkranz, et al., 2012):

(i) Many data mining techniques can easily and more efficiently handle huge amount of data.
(ii) Data mining does not make assumptions about the probability distribution of the data.
(iii) It aims to find patterns that can help to construct and formulate suitable hypotheses.

An overview of data mining techniques

- Scope of data mining techniques encompasses algorithms derived from "machine learning, pattern recognition, statistics and database theory".
- Types of machine learning:
  - supervised learning: the process of approximating a functional relationship to predict an output variable using input data.
  - unsupervised learning: the goal is to understand the underlying pattern, structure or distribution of the input data (no corresponding output variable).
  - deep structured learning: a new and emerging development of algorithms for multiple layers or hierarchical representations of data.

Types of data mining tasks

Generally four (4) main types we have identified useful for our purposes: regression, classification, association rule learning, and clustering.

- Labelled vs unlabelled data
- Labelled data has a specifically designated attribute and the aim is to use the given data to predict the value of that attribute for new data.
- Regression and classification work with labelled data and are considered supervised learning.
- Unlabelled data, on the other hand, does not have such a designated attribute.
- Association rule learning and clustering work with unlabelled data and are considered unsupervised learning.

Regression

Regression: widely popular, easy to interpret, and well understood

- actuarial applications: claims prediction, pricing, risk classification, claims reserving
  - Linear regression models and least squares
  - Given a vector of output **y** and an input vector of p features with design matrix **X**, our regression function assumes linear in the coefficient parameter β:

$$E(y|X) = f(X) = \beta_0 \sum_{j=1}^{p} X_j \beta_j$$

  - Least squares method assumes minimizing the residual sum of squares:

$$\widehat{\beta} = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 \right\},$$

  where $\text{RSS}(\beta) = ||\mathbf{y} - \mathbf{X}\beta||^2$.

Problems with least squares estimates

There are many situations when the LSE $\widehat{\beta}$ is not considered a very good estimator:

- large number of parameters (especially problematic for big data)
- the predictor variables are highly correlated
- produces unstable, highly variable estimates

One approach is to use shrinkage methods, or regularization, by introducing a penalty function.

These methods also help to reduce the variability of the prediction error.

Regularization or least squares penalty

- L_q penalty function:

$$\tilde{\beta} = \arg\min_{\beta} \left\{ ||\mathbf{y} - \mathbf{X}\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\},$$

  where λ is the regularization or penalty parameter.
- Special cases include:
  - LASSO (Least Absolute Squares and Selection Operator): $q = 1$
  - Ridge regression: $q = 2$
- Interpretation is to penalize unreasonable values of β.
- A variation is the use of elastic net penalty:

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta_j^2 + (1 - \alpha)|\beta_j| \right)$$

Classification and various methods

Classification: the problem deals with an output or attribute that is categorical or nominal, for example, assigning the attribute to one of a class $\{1, 2, \ldots, K\}$.

- actuarial applications: fraud detection, policy lapses/persistency, whether claim or not, risk classification
  - Regression for binary dependent variable:
    - Logistic regression: Cox (1958), linear predictors are linked to a logistic distribution.
    - Probit regression: Bliss (1934), linear predictors are linked to the normal distribution.
  - Classification using decision trees: involves the process of splitting points of attributes based on some specified criteria and applies pruning.
    - Pruning is the procedure of removing sections of trees to re-fine the classification for predictive accuracy.
    - CART algorithm: Breiman, et al. (1984), binary branches, pruning is done using bottom-up based on training data.

- continued

- Some methods based on Bayes Theorem:

$$\Pr(G = k|X = x)\frac{f_k(x)\pi_k}{\sum_{j=1}^{K} f_j(x)\pi_j}$$

where $f_j(x)$ are the class conditional densities of X given $G = k$ and π_j are the prior probability of being in class j, for $j = 1, \ldots, K$.

  - Linear discriminant analysis (LDA) assumes Gaussian class conditional densities
  - Naive Bayes: assumes very strong independence that each class density is a product of marginal densities.

- Support vector classifier: aim is to find a hyperplane that can separate the classes (linearly separable data).
- K-nearest negihborhood: classification based on some distance measure of nearness (e.g. Euclidean)

Association rule learning

- Agrawal, et al. (1993)
- Association rule learning is a method to find meaningful relations among incidence of events.
- Most useful for marketing products, but has large potential in actuarial applications e.g. fraud detection, policy lapse analysis.
- Start with a set of attributes $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ for which we can observe the occurrence:
  - Our dataset can be written as a set of T observations $D = \{d_1, d_2, \ldots, d_T\}$ where $d_t \subset \mathcal{A}$ for $t = 1, \ldots, T$.
  - Next define $i_t = (\mathbb{1}_{A_1 \subset d_t}, \ldots, \mathbb{1}_{A_n \subset d_t})$ so that $I = \{i_1, i_2, \ldots, i_T\}$ contains the same information with D as a dataframe.
  - Dataset is then stored and handled in the form of I.
  - If $X \subset \mathcal{A}$, then we call X is an itemset.

- continued

- An association rule is an if/then statement of the form:

$$X \Rightarrow Y, \quad \text{for itemsets } X \text{ and } Y.$$

- We call X the *precedent* and Y the *consequent*.
- For example in a basket of groceries, $\{\text{ground beef}, \text{onions}\} \Rightarrow \{\text{beer}\}$.

- Types of association rule:
  - single dimension: buys $\{\text{diaper}\} \Rightarrow$ buys $\{\text{baby foods}\}$
  - multiple dimensions: location (< 1), age $(< 22) \Rightarrow$ buys $\{\text{auto}\}$
  - hybrid dimension: career (< 1), buys $\{\text{car}\} \Rightarrow$ buys $\{\text{insurance}\}$

Measurement concepts

- Support of itemset X is the proportion of observations containing X in the whole dataset:

$$\text{supp}(X) = \frac{|\{d_t \subset \mathcal{A} | X \subset d_t\}|}{T}$$

It can be interpreted as the frequency to which you observe X in the dataset.

- Confidence indicates how often the rule between precedent and consequent is found among the given precedent:

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{|\{d_t \subset \mathcal{A} | X \subset d_t, Y \subset d_t\}|}{|\{d_t \subset \mathcal{A} | X \subset d_t\}|}.$$

It can be interpreted as the probability that a transaction with X also contains Y.

- continued

- Lift is the ratio of the observed support to that of the expected if X and Y are independent:

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)}$$

If X and Y are independent, there is no possible association rule that can be drawn. The lift is a measure of the degree to which there is dependent.

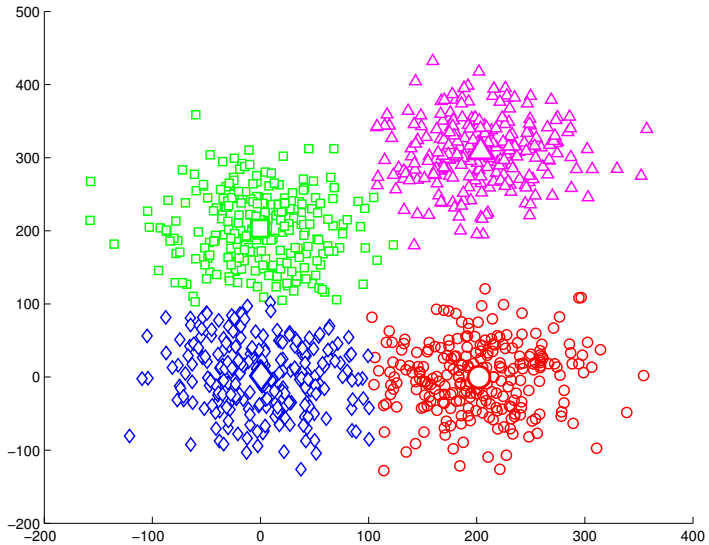
- Conviction is the ratio of how often the rule makes an incorrect prediction if X and Y are independent to that of the observed frequency of incorrect predictions:

$$\text{conv}(X \Rightarrow Y) = \frac{1 - \text{supp}(Y)}{1 - \text{conf}(X \Rightarrow Y)}$$

To illustrate, a conviction value of 1.35 indicates that you will be 35% correct that there is association than purely random.

Data clustering

- Data clustering, a form of unsupervised learning, is the process of dividing a group of data points into homogeneous groups or clusters.
- Data points in the same cluster share "similar" features while data points from different clusters are "dissimilar".



- Gan, et al. (2007), Gan (2011)
  - Has many applications in "big data" analytics: astronomy, medical science, marketing.
  - Has large potential in actuarial applications (general insurance; see Frees, et al. (2016), Ch. 6), valuation of large portfolios of VAs, mortality deviation for claims monotoring.
  - Many clustering algorithms have emerged in the last few decades.

Some common clustering methods

- Most clustering algorithms are formulated as an optimization problem.
- The well-known K-means algorithm performs clustering by minimizing the following objective function:

$$P(U, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} u_{il} ||x_i - z_l||^2,$$

where $\{x_1, \ldots, x_n\}$ is a dataset, k is the desired number of clusters, $U = (u_{il})_{n \times k}$ is an $n \times k$ partition matrix, $Z = \{z_1, \ldots, z_k\}$ is a set of cluster centers, and $|| \cdot ||$ is the L^2 norm or Euclidean distance.

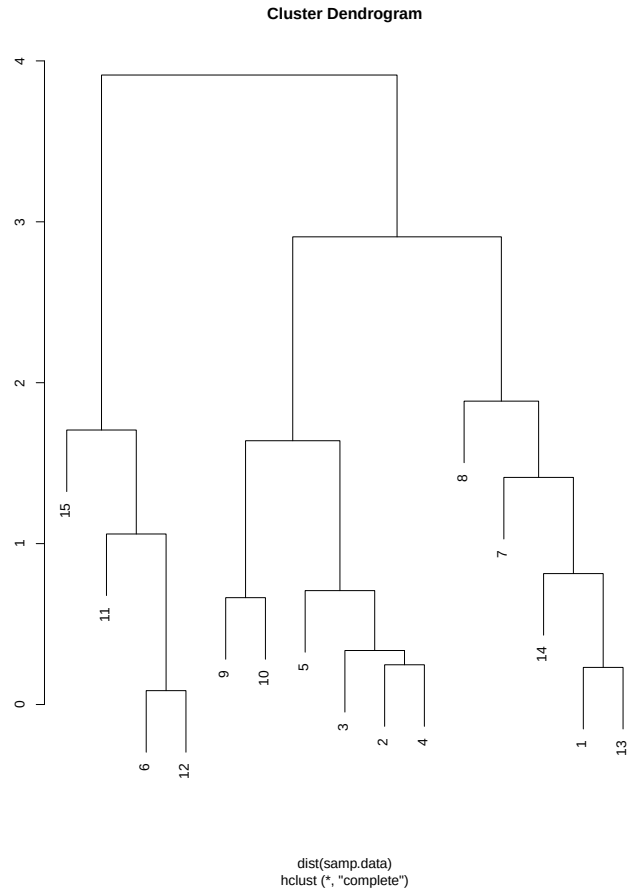
  - This results in an iterative process that partitions the dataset into a pre-specified number of clusters K.
  - At each iteration, each data point is assigned to its closest centroid (Euclidean distance).

- continued

- K-medoids is a variation proposed by Kaufman and Rousseeuw (1987).
- In contrast to using the means as center of cluster, K-medoids uses an actual data point in the cluster. This avoids the sensitivity of the K-means algorithm to extreme data points or outliers.
- Algorithm is also an iterative optimization process (steps are quoted directly from Hastie, et al. (2009)):
  - For a given cluster assignment C, find the observation in the cluster that minimizes the total distance to other points in that cluster. This gives us the current estimates of the cluster centers.
  - Given a current set of cluster centers, minimize the total error by assigning each observation ti the closest (current) cluster center.
  - Repeat the iterative process until the assignments do not change.

- continued

- Hierarchical clustering as the name implies, has the goal of building a hierarchy of clusters.
- This requires user to measure dissimilarity between groups of observations. Linkage criterion determines the measure of dissimilarity used: complete (max), single (min), average (mean), centroid (many variations).
- No need to pre-specify the number of clusters K. Two paradigms:
  - agglomerative (bottom up) clustering
  - divisive (top-down) clustering



- The *dendogram* is used to visualize the resulting clusters.

Concluding remarks

- Our goal is to survey data mining tasks associated with supervised and unsupervised learning, and explore their potentail use in actuarial problems for predictive analytics.
- We have only accomplished this mid-way in this presentation.
- We have lots of work to do yet:
  - Demonstrate their usefulness to actuarial problems: we have begun exploration of actuarial data to be used for each data mining tasks.
  - How the models can be used for predictive analytics and their performance, especially when compared to classical methods.
  - Short survey on the emerging development of a new class of machine learning algorithms called "deep learning".

Acknowledgment

We thank the financial support of the Society of Actuaries through our CAE grant on data mining.