**Advanced Unsupervised Learning methods applied to property-casualty databases**

*Application of Two Unsupervised Learning Techniques to Questionable Claims: PRIDIT and Random Forest*

***Louise A Francis, FCAS, MAAA***

**Abstract**

Predictive modeling can be divided into two major kinds of modeling, referred to as supervised and unsupervised learning, distinguished primarily by the presence or absence of dependent/target variable data in the data set used for modeling. Supervised learning approaches probably account for the majority of modeling analyses. This paper will focus on two infrequently used unsupervised approaches: PRIDIT and Random Forest clustering.

Unsupervised learning is a kind of analysis where there is no explicit dependent variable. Examples of unsupervised learning in insurance include modeling of questionable claims (foe some action such as referral to a Special Investigation Unit) and the construction of territories by grouping together records that are geographically “close” to each other. Databases used for detecting questionable claims analysis often do not contain a fraud indicator as a dependent variable. Unsupervised learning methods are often used to address this limitation. The PRIDIT (Principal Components of RIDITS) and Random Forest (a tree based data-mining method) unsupervised learning methods will be introduced. We will apply the methods to an automobile insurance database to model questionable[[1]](#footnote-1) claims.

A simulated database containing features observed in actual questionable claims data was developed for this research based on actual data. The database is available from the author.

Introduction

An introduction to unsupervi**s**ed learning techniques as applied to insurance problems is provided by Francis (2013) as part of a predictive modeling text intended to introduce actuaries to predictive modeling analytic techniques. As an introductory work, it focused on two classical approaches: principal components and clustering. Both are standard statistical methods that have been in use for many decades and are well known to statisticians. The classical approaches have been augmented by many other unsupervised learning methods such as kohonen neural networks, association rules and link analysis. The two methods we feature here, PRIDITS and Random Forest clustering are less well known and less widely used. Brockett et al. (2003) introduced the application of PRIDITs to the detection of questionable claims in insurance. Brieman and Cutler incorporated and unsupervised learning capability into their open source software for preforming Random Forest. Shi and Horvath (2006) provided an introduction to how the method works, along with a tutorial for implementing Random Forest clustering.

Unsupervised Learning

“Unsupervised learning”, a term coined by artificial intelligence professionals, does not involve dependent variables and predictors. Common unsupervised learning methods include cluster analysis and principal components analysis. Francis (2014) provides an introduction to unsupervised learning for actuaries along with examples of implementation in R.

Why develop a model that does not have a dependent variable? To motivate an understanding of applying unsupervised learning in insurance, we use the questionable claims example. A common problem in claims data is the absence of a dependent variable. That is the claims data that are used to construct questionable claims models do not clearly label the records as to whether a questionable claim is suspected. Sometimes that data contain surrogates such as whether an independent medical exam was ordered or whether a referral was made to a special investigation unit (Derrig and Francis, 2008). Sometimes unsupervised learning is used to address this challenge. Two overall approaches are available:

* Use unsupervised learning methods to construct an index or score from the variables in the file that have been found related to questionable claims. Brockett et al. (2003) showed how the PRIDIT technique can be used to perform such a task. PRIDIT will be one of the methods used in this paper.
* Use a clustering type of procedure to group like claim records together. Then examine the clusters for those that appear to be groups of suspected questionable claims. Francis (2014) showed a simple application of clustering to classify claims as into legitimate vs. suspected questionable claims. Derrig and Ostasurski (1995) showed how fuzzy clustering could be used to identify suspected questionable claims. The second method used in this paper, Random Forests, will be compared to clustering.

If a variable related to questionable claims is in fact available in a dataset, as is the case with the PIP claims data (both original data and simulated data), the unsupervised learning methods can be validated.

Simulated Automobile PIP Questionable Claims Data and the Fraud Issue

Francis (2003, 2006), Viaene (2002) and Brockett et al. (2003) utilized a Massachusetts Automobiles Insurers Bureau research data set collected in 1993. The data set was from a closed claim personal automobile industry PIP data set. The data contained data considered useful in the prediction of questionable claims. This data was the basis of our simulated data used in this paper. We have used the description of the data as well as some of the published statistical features to simulate automobile PIP claims data. The use of simulated data enables us to make the data available to readers of this paper[[2]](#footnote-2). Note that while we incorporate variables found to be predictive in prior research, our variables do not have the exact same correlations, importance rankings or other patterns as those in the original data. As the original data had over one hundred variables in it, only a subset have been considered for inclusion in our simulated data.

Note that in this paper a distinction is made between questionable claims and fraud. Derrig (2013) in a presentation to the International Association of Special Investigative Units went into some detail about the difference. Historically, fraud has often been characterized as “soft fraud” and “hard fraud”. Soft fraud would include opportunistic activities, such things as claim build-up (perhaps in order to exceed the state’s PIP tort threshold) and up-coding of medical procedures in order to receive higher reimbursement. Soft fraud has sometimes been referred to as “abuse”. Hard fraud is planned and deliberate includes staged accidents and filing a claim (where the claimant may have been paid to file the claim on behalf of a fraud ring) when no accident in fact occurred[[3]](#footnote-3). According to Derrig, from a legal perspective for an activity to qualify as fraud it must meet the following four requirements: (1) it is a clear and willful act; (2) it is proscribed by law, (3) in order to obtain money or value (4) under false pretense. Abuse or “soft fraud” fails to meet at least one of these principals. Thus, in many jurisdictions activities that are highly problematic or abusive are not illegal and therefore do not meet this definition of fraud. Derrig believes that the word “fraud” is ambiguous and should be reserved for criminal fraud. Although a rule of thumb often cited that 10% of insurance claims costs are from fraudulent[[4]](#footnote-4) claims, Derrig asserts, based on statistics compiled by the Insurance Fraud Bureau of Massachusetts ,that the percentage of criminal fraud is far lower, perhaps a couple of percent. As a result of this, Derrig, and many other insurance professionals use the term “questionable claims” rather than “fraud”.

Derrig advocated the use of predictive analytics to the questionable claims problem. Derrig and others (see Derrig and Weisberg 1995) participated in assembling a sample of claims from the Automobile Insurers Bureau of Massachusetts (AIB) from the 1993 accident year in order to study the effectiveness of protocols for handling suspicious claims and to investigate the possible use of analytics as an aid to handling claims. The sample data contained two kinds of data: claims file variables and red flag variables. The claims file variables represent the typical data recorded on each claim by insurance companies. Thus, the claims data contains numeric (number of providers, number of treatments, report lag) as well as categorical variables (injury type, provider type, whether an attorney is involved, whether claimant was treated in the emergency room).

The “red flag” variables were particular to the study and are not variables typically contained in insurance claims data. These variables are subjective assessments of characteristics of the claim that are believed to be related to the likelihood it is/ is not a legitimate claim. The red flag variables had several categories such as variables related to the accident, variables related to the injury, variables related to the insured and variables related to the treatment. An example is the variable labeled “ACC09”, an accident category variable denoting whether the claims adjustor felt there was no plausible explanation for the accident.

Based on concepts and relationships observed in the AIB data, a simulated database was created. The simulated data for this paper has 1,500 records. The original AIB data has 1,400 records. The database contained simulated predictor variables similar (i.e., in variable name, and direction of the relationship to the target) to actual variables used in previous research from the original database, as well as a simulated target variable. The original data contained several potential dependent variables. The simulated data contains only one, an indicator variable that indicated whether the claim is believed to be questionable. An advantage of this simulated data is that it can be free shared with others and can be used to do research on the methods in this paper. Table 1 shows the red flag variables in the simulated data. Table 2 shows the claim file variables.

**Table 1 Red Flag Variables**

|  |  |
| --- | --- |
| **Variable** | **Label** |
| Inj01 | Injury consisted of strain or sprain only |
| Inj02 | No objective evidence of injury |
| Inj06 | Non-emergency treatment was delayed |
| Ins06 | Was difficult to contact/uncooperative |
| Acc04 | Single vehicle accident |
| Acc09 | No plausible explanation for accident |
| Acc10 | Claimant in old, low valued vehicle |
| Acc15 | Very minor impact collision |
| Acc19 | Insured felt set up, denied fault |
| Clt07 | Was one of three or more claimants in vehicle |

**Table 2 Claim File Variables**

|  |  |
| --- | --- |
| **Variable** | **Label** |
| legalrep | claimant represented by a lawyer |
| sprain | back or neck sprain |
| chiropt | Chiropractor/physical therapist used |
| emtreat | received emergency room treatment |
| police | police report filed |
| prior | history of prior claims |
| NumProv | Number of health care providers |
| NumTreat | Number of treatments |
| RptLag | Lag from accident to date of report |
| TrtLag | Lag from accident to date of 1st treatment |
| PolLag | Lag from policy inception to accident |
| Thresh | damages exceed threshold |
| Fault | percent policyholder fault |
| ambulcost | cost of ambulance |

The original AIB data contained two overall assessments from separate specialists as to whether the claim was legitimate or whether it was believed to have some degree of suspicion. The variables were in the form of scores, which ranged from 0 to 10 and 1 to 5. Note that these were subjective assessments, as the actual classification of the claims was unknown. For both variables the lowest category (0 and 1 respectively) denoted a legitimate claim[[5]](#footnote-5). In the simulated data, the binary variable “suspicion” indicates the claim specialist examining the file suspected a questionable claim[[6]](#footnote-6). The variable is coded 1 for no suspicion of fraud and 2 for suspicion of fraud. Approximately 1/3 of the records have a coding of 2. This variable will be used in assessing the unsupervised learning methods but not in building the unsupervised models

.

The Questionable Claims Dependent Variable Problem

Insurance claims data typically do not contain a variable indicating whether the claim is considered questionable. That is, even if the claims adjuster or other insurance professional considers the claim suspicious, there is no record in the claims data of that suspicion. Certain surrogate variables may capture information on some claims. For instance, many claims database contain a variable indicating that a referral is made to a special investigation unit (SIU). However, only a small percentage of claims receive an SIU referral (as these frequently represent claims suspected to be criminally fraudulent), so the absence of a referral does not mean that the claim was deemed legitimate. The absence of a target variable in most claims databases suggests that an unsupervised learning approach could be very helpful. For instance, if unsupervised learning could be applied to the features in the data to develop a score related to whether the claim is questionable, the score could be used to classify claims for further handling, such as referral to an SIU. The PRIDIT method is an approach to computing such a score from claim predictor variables, when a target variable is not present.

The PRIDIT Method

PRIDIT is an acronym for Principal Components of RIDITS. RIDITS are a percentile based statistic. The RIDIT transformation is generally applied to variables whose values can be considered in some sense to be ordered. These might be answers to a survey (i.e., disagree, neutral, agree, etc.), but the variables might also be binary categorical variables such as the red flag variables in our claims data, where one of the values on the variable is believed to be related to suspicion of a questionable claim and the other to a likely legitimate claim.

Bross (1958) in his paper introducing RIDITS says that in a number of his studies it could simplify complex data and make it possible to answer some of the investigator’s questions (Boss, 1958, p19). The RIDIT statistic is considered distribution free. Bross states the term “RIDIT” was selected to have similarity to “probit” and “logit”, two common transformations of categorical data, and the first three letters stand for *R*elative to an *I*dentified *D*istribution. It is a probability transformation based on the empirical distribution of data. Boss also views it as a way to assign a weight to the categories of ordered data. Boss notes that the RIDIT may be assigned based on a “base group”, say healthy people in a study of a medical intervention. In the example in his paper, he used a 10% sample of his car accident dataset to calculate the ridits.

For an ordered categorical variable X, where the values of X can be numbers (such as 1, 2, etc.) or qualitative values (such as low, medium, high), first compute the proportion of records in each category. Then compute the cumulative proportion for each value of X (going from low values of X to high values of X). TheRIDIT formula for the RIDIT based on these empirically based probabilities is:

1. 

In Table 3 we show the calculation of a RIDIT for a theoretical injury severity variable

**Table 3 Calculation of a Ridit**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Injury Severity** | **Count** |  **Cumulative Count** | **Probability** | **Cumulative Probability** | **RIDIT** |
| Low | 300 | 300 | 0.3 | 0.3 | 0.15 |
| Medium | 600 | 900 | 0.6 | 0.9 | 0.6 |
| High | 100 | 1000 | 0.1 | 1 | 0.95 |
| Total | 1000 |  | 1 |  |  |

A different version of the formula is given by Brockett et al. (2003):

(2)

Note that Boss’s RIDIT ranges from zero to one, while the Brocket et al. RIDIT can range from -1 to 1. Although these two definitions of the RIDIT score appear to be somewhat different, they are actually behave similarly and under certain assumptions the Brockett et al. RIDIT is a linear transformation of the RIDIT as defined by Boss.[[7]](#footnote-7) If one assumes that ½ of the category P(X = Xi) belongs to P(X< Xi) and one half to P(X > Xi) then the transformation 2 \* RIDITBoss – 1 produces the Brockett et al. RIDIT.

PRIDITs involve performing a principal components or factor analysis on the RIDITS. This approach is distinguished from classical principal components in that the principal components procedure is applied to a transformation of the original data. Brocket and Derrig (2003) introduced this approach to insurance professionals and applied the technique to investigating questionable claims where it was found to be an effective approach to identify questionable claims. The data used in the Brockett et al. study was conceptually similar to the dataset we use in this paper. However, the questionable claims data in this paper is simulated, not actual data. As noted by Brockett et al. (2003), a useful feature of the PRIDIT method is that each variable in the PRIDIT score is weighted according to its importance. That is when developing a score to classify claims, one wants to give greater weight to the variables most related to whether the claim is questionable.

Processing the Questionable Claims data for PRIDIT analysis

All categorical variables were already in binary form, where computing the RIDIT is straightforward. In general, variables that were originally had multiple categories were turned into one or more binary variables (such as sprain versus all other injuries). The numerical variables were binned into no more than 10 levels. In general. for variables with no zero mass point, such as report lag, were split into even bins (in the case of report lag 5 bins) based on quantiles. Other variables with a zero mass point such as ambulance cost (many claimants did not use an ambulance) were binned judgmentally, with the first bin (often with over 50% of the records) containing the zero records.

Computing RIDITS and PRIDITS

In general statistical software for computing the RIDIT transform of a variable is not widely available. According to Peter Flom of Statistic Analysis Consulting[[8]](#footnote-8) SAS’s Proc FREQ can do RIDIT analysis. In addition, R has a RIDIT library. However, we were unable to adapt that function to our application requiring outputting an individual RIDIT scores to claims records. Therefore we developed R code specifically for computing RIDITS and assigning the RIDITS to each record[[9]](#footnote-9). The R code is available from the author. Once the RIDITS have been calculated, principal components/factor analysis can be performed in virtually any statistical software package. We have used both SPSS (factor analysis) and the princomp and factanal functions from R. Venables and Ripley (1999) show examples of the use of the R princomp function.

PRIDIT Results for Simulated PIP Data

A PRIDIT analysis was performed on the 10 red flag variables and fourteen claim file variables. The scree plot below shows that the first few factors explain a large percentage of the variability of the predictor variables (i.e., of the RIDITS). The first component explains about 25% of the variability.

**Figure 1**



The magnitude of the loading of the RIDITS on the components informs us as to the relative importance of each variable. Table 4 displays the loadings on the first two factors for the top 10 (by absolute value) variables.

**Table 4 Component Loading**



From these loadings we can tell that injury (i.e. sprain)and treatment variables (Number of providers, use of chiropractor) have a high loading (in absolute value terms) on the first component, along with whether the claimant has legal representation.

How good is the PRIDIT score?

Brocket et al. (2003) used the suspicion scores assigned by experts in a confirmatory analysis of their PRIDIT results. In the original AIB data the scores were subjectively assigned. The PRIDIT method as presented by Brockett and Derrig (2003) used only the first component. Brockett et al. noted after regressing the scores on the claim file variables, that the apparent weights assigned by claims experts did not always agree with that assigned by PRIDIT. In their view the differences highlight a weakness of subjective scores. That is that there are inconsistencies between different experts making an evaluation with the same data, and the experts may under or overweight some variables when making a judgment involving many variables. They concluded that a value of the PRIDIT procedure is its ability to objectively determine the weights assigned to each variable based on the evidence in the data.

In our simulated dataset, the target variable has not been subjectively assigned, but has been simulated from an underlying distribution, where on average about 30% of the claims are expected to be in some way questionable. In our assessment the PRIDIT is taken to be the first principal component/factor[[10]](#footnote-10) only. A quick analysis indicated that other components, especially the second also had predictive value, though the value of second and higher components is not explored further in this paper. Table 5 below displays the means of the PRIDIT score for each of the fraud categories as measured by the suspicion variable.

**Table 5/Figure 2: Means of PRIDIT Score with Error Bar Graph**





The statistics indicates a strong relationship between the PRIDIT score and propensity for questionable claims. Note that the PRIDIT score appears to be negatively correlated with the probability of being a questionable claim. The nature of the relationship can be discerned by evaluating the relationship of the score to some of the variables. For instance, the average scores for records that are positive for the sprain injury category is negative as is the average score for records that are positive for chiropractor/physical therapist use. Moreover ambulance cost is often negatively related to suspicion, as claimants truly injured in an accident are transported by ambulance to a hospital, but not questionable claimants. Both of these have been shown in other studies to be positively related to suspicious claims (Francis 2003, Francis 2012[[11]](#footnote-11)).

A common tool in evaluating the effectiveness of a numeric predictor for predicting a binary target variable is the Receiver Operating Characteristics (ROC) curve. See Gareth et al. (2013) for a discussion of ROC curves. A variable whose curve rises well above the diagonal line and exhibits an area under the ROC curve (AUROC) substantially in excess of 0.5 is considered a good predictor. Below is the ROC curve from the PRIDIT analysis. The ROC curve indicates a strong relationship between the PRIDIT score and the “true” suspicion category.

**Figure 3: ROC Curve for Suspicion Category using PRIDIT as Predictor**





Trees and Random Forests

In the past decade, computationally intensive techniques collectively known as *data mining* have gained popularity for explanatory and predictive applications in business. Many of the techniques, such as neural network analysis, have their roots in the artificial intelligence discipline. One of the most popular of the data mining tools, decision trees, originated in the statistics discipline, although an implementation of trees or classification and regression trees (C&RT) known as C4.5 was independently developed by artificial intelligence researchers. The classic book on trees by Brieman et al. (1984) provided an introduction to decision trees that is still considered a standard resource on the topic. Trees use a recursive partitioning approach. DeVille (2006) provides an easy to understand introduction to a variety of decision tree methods, however, the illustrations do not use insurance data. Derrig and Francis (2008) provide an introduction to tree method that use insurance data and are therefore appropriate for an actuarial audience.

Decision trees are run in two modes: classification and regression depending on whether the dependent variable is categorical or numeric. A different loss function is optimized for each of the two modes. Details of the loss functions are covered in deVille (2006) and Francis and Derrig (2008)

Many tree models can output an importance ranking of the predictor variables in the model. We will be utilizing this feature later when we compare Random Forest (RF) clustering against classical clustering. Two reasons for the popularity of decision-tree techniques are (1) the procedures are relatively straightforward to understand and explain, and (2) the procedures address a number of data complexities, such as nonlinearities and interactions, that commonly occur in real data. In addition, software for implementing the technique, including both free open source as well as commercial implementations, has been available for many years. In addition, the tree method is believed to be robust to outliers among predictor variables, as the cut point for the trees (for numeric variables) is based on ordinal relationships. Once a tree model has been fit, the relationships can be displayed graphically as an upside down tree. An example of the output of tree used to model the dependent variable Suspicion (denoting a legitimate or suspicious claim) is shown in Figure 4.

**Figure 4: Example of Decision Tree**



Ensemble models are composite tree models. A series of trees is fit and each tree improves the overall fit of the model. The ensemble model’s prediction is a weighted average of the single tree predictions. In the data mining literature two common ensemble techniques are referred to as “boosting” (Hastie, Tibshirani, and Friedman 2001; Freidman 1999) and Random Forests (Brieman, 2001). As boosting is not the focus of this paper it will not be discussed further. Random forest uses the ensemble method of “bagging”. Bagging is an ensemble approach based on resampling or bootstrapping. Bagging is an acronym for “bootstrap aggregation” (Hastie et al. 2001). Bagging uses many random samples of records in the data to fit many trees. For instance, an analyst may decide to take 50% of the data as a training set each time a model is fit. Under bagging, 100 or more models may be fit, each one to a different 50% sample. The trees fit are unpruned and may be large trees with more than 100 terminal nodes. By averaging the predictions of a number of bootstrap samples, typically using a simple average of all the models fit, bagging reduces the prediction variance.[[12]](#footnote-12) The implementation of bagging used in this paper is known as Random Forest. Brieman (2001) points out that using different variables, as well as different records in the different trees in the Random Forest ensemble, seem to reduce the correlation between the different models fit and improve the accuracy of the overall prediction.

Unsupervised Learning with Random Forest

Shi and Horvath (2006) provide an excellent overview of using Random Forest for unsupervised learning. When run in unsupervised learning mode, the Random Forest procedure can produce a proximity matrix. The proximity matrix is a measure of how similar one record is to another record. The measure is based on how often the records being compared are in the same terminal node of stand-alone trees. Keep in mind the Random Forest method produces an ensemble of single trees which are weighted together to produce a final prediction for each record. In computing a proximity measure each tree gets a vote. The total number of times each of two records being compared are in the same terminal node divided by the total number of tree is their proximity between the two records. This number will vary between zero and one.

The proximity statistic just described can be computed whenever Random Forest is run as a supervised learning tool. To compute a proximity matrix in an unsupervised learning application a dependent variable is artificially created. This pseudo dependent variable is created as an artifact of creating a simulated or “synthetic” database that does not contain the structure, i.e. the dependencies and interactions, of the original data. Shi and Horvath mention two methods of simulating the synthetic data. The most common approach referred to by Shi and Horvath (2006) as “Addcl1” in their paper and describe as sampling from the product of marginal distributions. It is equivalent to independent sampling from each variable. This is as opposed to randomly sampling from each record where all values of all variables are used once the record is randomly selected. With independent sampling only one value for one variable of one record is selected with each draw. Thus the dependencies are broken and the “synthetic” data can be contrasted with the actual data in a supervised learning model. A binary variable is created that indicates which data the record is from (actual or synthetic), and a Random Forest is run as a supervised learning model with the binary variable as the dependent variable. From the Random Forest ensemble model a proximity matrix can be computed as described above. The proximity between any two records (I and j) can then be transformed into a dissimilarity using the following formula per Shi and Horvath:

(3) 

The dissimilarity matrix can then be used as input to a standard clustering procedure. In our examples, we use the pam (partition against medoids) function of the R cluster library, per the examples and tutorial of Shi and Horvath.

Software for Random Forest Computation

For the analysis in this paper, the R randomForest library was used. In addition to the randomForest procedure, a clustering procedure is needed. In this paper the R cluster library was used. An introduction to clustering and the R cluster library is provided by Francis (2014).

Application of Random Forest Clustering to Suspicious Claims Data

Random Forest clustering was applied to the simulated questionable claims data predictor variables. When a dependent variable is not specified the R random Forest function is run in unsupervised mode. The approach was:

* The claim file variables and the red flag variables are input into the R randomForest function from the randomForest library
* The number of trees was chosen to be 500 reflecting the recommendations of Shi and Horvath that many trees be fit
* A proximity matrix was output
* The procedure was repeated a second time to produce a second proximity matrix was computed, per the recommendation of Shi and Horvath that multiple forests be run
* The two proximity matrices were averaged
* The dissimilarity matrix was computed from the average proximity matrix
* The dissimilarity matrix was supplied to the pam clustering procedure to perform kmeans clustering
* K= 2,3 and 4 clusters were computed and for each scenario the each record was assigned to one of the clusters
* For comparison the pam function was used to run an ordinary kmeans (i.e., not using RF) clustering on the predictor variables. The dissimilarities used in the clustering used the Euclidean distance measure.

The “true” value for the suspicion variable was used in validating the clusters. An objective is to determine if one or more of the clusters created by the RF clustering and the Euclidean clustering seemed to have a high representation of suspicious claims. The two techniques of evaluation we will use are:

* The chi squared statistic

This classic statistic is used to evaluate when there is a statistically significant relationship between cross classified variables. The higher the value of the statistic, the stronger the relationship between two categorical variables.

* The tree importance ranking of the different variables

Once clustering has been performed, the various clusters can be used as predictors in a tree model as seen in Figure 5, where suspicion is the dependent variable and the clusters are the predictors. The model will rank each variable in importance, assigning the highest statistic (and ran as first) to the variable which contributes the most to minimizing the loss function. Due to the small number of predictors in this test, a simple tree rather than a more complicated Random Forest model was used. The rpart function from R, which is a very common function from R used for tree modeling, was used.

The statistics were examines to determine which K (i.e., how many clusters) seemed to perform the best separation of records and which clustering method (RF clustering or Euclidean clustering) performed best. The following shows the chi square statistics:

**Table 6: Chi Square Statistic for Random Forest and Euclidean Clusters**

|  |  |  |
| --- | --- | --- |
| **K** | **Random Forest** | **Euclidean** |
| 2 | 183 | 254 |
| 3 | 604 | 191 |
| 4 | **715** | 254 |

The Chi Square statistic suggests that the Random Forest clusters for k= 4 groups does the best job of segmenting claims into groups based on the probability of suspicious claims. The table of the proportion of suspicious claims with each cluster is shown for the Random Forest and Euclidean clusters with k=4 groups:

**Figure 5: Tree of Suspicion of Questionable Claims and Cluster[[13]](#footnote-13)**



**Table 7: Proportion of Suspicious Claims by Cluster**

|  |  |  |
| --- | --- | --- |
|  | **Random Forest** | **Euclidean** |
| **Group** | **% Suspicious** | **N** | **% Suspicious** | **N** |
| 1 | 24% | 501 | **54%** | 536 |
| 2 | 2% | 389 | 5% | 365 |
| 3 | **91%** | 317 | 26% | 390 |
| 4 | 17% | 293 | 25% | 209 |

The statistics for the Random forest clusters indicate that a high suspicion cluster has 91 percent of claims that are suspicious and a low suspicion cluster has 2% of claims that are suspicious. Two other random forest clusters have mean somewhat modest mean percentages (17% and 24%). This compares to the Euclidean clustering where the high suspicion cluster has 54% suspicious claims and the low suspicion cluster has 5% suspicious claims and two moderate clusters have about 25% suspicious claims.

In actual analyses using unsupervised learning, our data would not contain the independent variable so we would have to use the other variables in the data to subjectively label clusters as low, medium and high suspicions level groups. For instance, we could examine a table of some of the key statistics defining the clusters.

**Table 8: Descriptive Mean Statistics for Random Forest Clusters**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| RF Four Group Cluster | Sprain | Chiro/pt | NumProv | NumTreat | TrtLag | Thresh | ambulcost |
| 1 | .83 | .91 | 2.11 | 2.92 | 1.79 | .84 | 62.81 |
| 2 | .07 | .31 | 3.32 | 4.52 | 1.61 | .81 | 270.54 |
| 3 | .88 | .93 | 2.62 | 3.62 | 14.95 | .31 | 203.58 |
| 4 | .09 | .30 | 5.98 | 8.72 | 4.55 | .32 | 273.78 |
| Total | .50 | .64 | 3.29 | 4.62 | 5.06 | .62 | 187.64 |

 Table 8 indicates that the high suspicion (group 3) versus the low suspicion group (group 2) is characterized by a high proportion of claims involving sprains and the use of a chiropractor/physical therapist, a significantly longer treatment lag and lower proportion of claims meeting a tort threshold. Also the low suspicion group has higher average ambulance costs (presumably because the presence of an obvious significant injury leads to the use of an ambulance). The average number of treatments is higher for the low suspicion group, but these differences are not that large.

A simple Tree model was run separately for the RF clusters and the Euclidean clusters and then for both. The models were used to determine the importance ranking of each of the predictors. That is, if the cluster variable for k=2 was ranked number 1 in importance, we have an indication that k=2 is the best selection for classifying claims into legitimate/questionable classes. For the RF clustering, the 4-group cluster was ranked most important, while for Euclidean clustering the 2-group cluster was ranked most important. These ranking agree with those produced by the Chi-square statistic. When both the RF clusters and the Euclidean clusters were used as predictors in the tree model, the RF 4-group cluster was ranked #1 and the RF 2-group cluster was ranked #. The Euclidean 4-group cluster ranked fourth in importance, behind all of the RF clusters.

**Table 9: Ranking of Random Forest Clusters in Predicting Suspicion**



**Table 10: Ranking of Random Forest and Euclidean Clusters in Predicting Suspicion**



The importance rankings were also used to compare the PRIDIT score to the clusters. The tree ranked the RF4-group cluster as most important in predicting questionable claims, followed by the PRIDIT score. In general, these tests indicate that the two unsupervised techniques featured in this paper outperform classical clustering in identifying suspicious claims. Note that in a previous analysis (Francis, 2012), the PRIDIT score ranked higher than the RF clusters.

**Figure 6: Ranking of PRIDIT, RF Clusters and Euclidean Clusters in Predicting Suspicion**



It should be noted that the Random Forest procedure can rank variables in importance to the model. This ranking can be contrasted with the loadings from the principal components procedure used in the PRIDITs. The comparison indicates significant differences between the two procedures in the importance of variables.

**Table 11: Ranking of Top 10 Variables in Random Forest Clustering**



Some Findings from the Brockett et al. Study

Brockett et al. (2003) investigated a number of data and model characteristics in their 2005 study of the AIB research sample including:

* How consistent were insurance expert evaluations of claims
* How consistent was the PRIDIT score with adjuster evaluations

The AIB data contained two different evaluations of the likelihood the claim was questionable (1) an assessment score (on a scale of 0 – 10) from a claims adjuster reflecting the likelihood the claim was questionable and (2) an assessment score from a fraud expert from the Insurers Fraud Bureau of Massachusetts

Brockett et al. performed a regression analysis using each of the two assessment scores as a dependent variable. They found a relatively high spearman (nonparametric) correlation (0.78) between the PRIDIT score and the prediction from the regression on the adjuster’s score, but a lower correlation (0.60) between the PRIDIT score and the actual adjuster’s score. A similar pattern was observed between the PRIDIT score and the regression prediction from the investigator’s score (0.64) and the actual investigator’s score (0.49). Brockett et al. suggest that the adjuster and investigator regression scores are more objective evaluations as people have difficulty weighing together many factors at one time. The AIB data contains 65 red flag indicators that were used by the adjusters and investigators. Determining the weight to give to each red flag variable in an overall score can be a challenge, and different adjusters/investigators will assign different weights. Brockett et al. note that even though individuals may have difficulty weighing together individual variables the subjective assessments of insurance experts are needed to score the individual red flag variables. They also suggest that the lower correlation between the PRIDIT score and both the regression prediction and actual value from the investigators score is because the investigators focus more on criminal fraud as opposed to other kinds of abuse. They note some inconsistency between the adjusters’ and the investigators’ actual scores, i.e., a correlation between 0.56 and 0.60. Brockett et al. also suggest that when assessing claims, the adjusters focus on information that will help them adjust the claim while the investigators focus on information related to legal fraud, thus their evaluation framework is different.

The Brockett et al. study findings appear to support (1) the value of analytic models for evaluating questionable claims, whether supervised or unsupervised, and (2) the PRIDIT method as applied to the AIB data appeared to produce a score that was consistent with scores obtained from supervised learning modes.

In this study the unsupervised learning methods have only been tested for consistency against a known (simulated) target variable. The data used for this paper, because it is simulated, does not contain a “true” subjective assessment of a claim, therefore comparisons like those of Brockett et al. cannot be performed.

A purpose of this paper was to provide an instructional example of the application of advanced unsupervised learning methods. Therefore a dataset was used that can be shared with other researchers but does not contain all of the features of the original data upon which it is modeled.

Summary

Two unsupervised learning methods, PRIDIT and Random Forest Clustering were applied to insurance data. One objective of unsupervised learning is to score records for characteristics of interest (such as suspicion of submitting questionable claims) in the absence of an identifiable dependent variable in the data. The illustration in this paper demonstrated the possible usefulness of PRIDIT, a principal components technique applied to the RIDIT transformation of variables and RF clustering an ensemble tree based method that can be used for finding groups in the data. The PRIDIT score could be used to classify claims for further handling. Using RF clustering, one or more of the groups may be identified for follow-up action such as request for an independent medical exam (to validate the claim) or referral to a special investigation unit (for investigation of abuse or fraud).

The “true” suspicion value was used in order to assess the unsupervised methods in this paper. On the simulated questionable claims data the R rpart library was used to build a tree model to predict questionable claims, with the PRIDIT score, the RF Clusters and classical Euclidean clusters as predictors. One of the outputs of the tree model is a ranking of the variables in importance in predicting the target. Both RF clustering and the PRIDIT method performed well in the testing and both outperformed classical Euclidean clustering in identifying questionable claims. Similar results were found by Francis (2012) when applied to the actual 1993 AIB data, rather than simulated data.

A useful feature of the PRIDIT method is that it produces a single score that can be used to sort claims on questionableness (or on some other feature of interest to a business). The Random Forest clustering method does not have this feature, and further analysis using descriptive statistics from the variables used for clustering is required to identify the clusters that contain questionable claims.

References

1. Aldender M. and Bscorelashfield R, *Cluster Analysis*, Sage Publications,1984
2. Brieman L, Cutler A, Liaw, Weiner, Package randomForest, [www.r-project.org](http://www.r-project.org)
3. Brieman L., Friedman J. H., Olshen R and Stone C., Classification and Regression Trees, Chapman and Hall/CRC, 1984
4. Breiman, L, “Random Forests”, Machine Learning 45 (1) 5-32
5. Brockett, Patrick L., Derrig, Richard A. Golden, Linda L. Levine, Arnold
Alpert, Mark, “Fraud Classification Using Principal Component Analysis of RIDITs “,
*Journal of Risk and Insurance, The American Risk and Insurance Association - Mount Vernon, New York* 2003: 69:3 341-371
6. Brockett, P and Levine A, “On A Characterization of RIDITS”, Annals of Statistics, 1977, pp 1245 – 1248
7. Bross, I, “How to Use Ridit Analysis”, Biometrics, , 14 (1):18-38, 1958 – available from JSTOR
8. Coaley, Keith, An Introduction to Psychological Assessment and Psychometrics, Sage, 2010
9. De Ville B., *Decision Trees for Business Intelligence and Data Mining*, SAS Institute, 2006
10. Derrig R. and Weisberg H., “A Report on the AIB Study of 1993 PIP Claims, Part 1, Identification and Investigation of Suspicious Claims”, Automobile Insurers Bureau of Massachusetts, 1995
11. Derrig, R., “Using Predictive Analytics to Uncover Questionable Claims”, presentation to the International Association of Special Investigative Units, Sept 9, 2013
12. Francis, Louise, “Martian Chronicles: Is MARS Better than Neural Networks”, 2003 CAS Winter Forum, “Unsupervised Learning Applied to Insurance Data”, 2012 Salford Data Mining Conference
13. Francis, Louise, “Review of PRIDIT”. Presentation at 2006 CAS Ratemaking Seminar
14. Francis L, “Unsupervised Learning Applications in Property Casualty Insurance”, presentation at Salford Systems Data Mining Conference, 2012
15. Francis, L. draft of “Unsupervised Learning”, Chapter 12 in *Predictive Modeling Applications in Actuarial Science*, 2013, book to be published in 2014
16. Friedma, J. “Stochastic Gradient Boosting”, <http://statweb.stanford.edu/~jhf/ftp/stobst.pdf>, 1999
17. Friedman, Jerome, Hastie, Trevor and Tibsharani Robert, *The Elements of Statistical Learning*, Springer, 2001
18. Gareth J., Witten D. Hastie T. Tibshirini R., *An Introduction to Statistical Learning*, Springer, 2013
19. Insurance Information Institute information on fraud, http://www.iii.org/facts\_statistics/fraud.html
20. Kaufman Leonard and Rousseeuw, Peter, *Finding Groups in Data* John Wiley and ns, (1990)
21. Kim Jae-on and Mueller, Factor Analysis: Statistical Methods and Practical Issues, Sage University Press, 1978
22. Maechler, Martin, ,”Package ‘cluster’” downloaded from CRAN web site (see [www.r-project.org](http://www.r-project.org)), Feb 2012
23. Shi, T and Horvath S., “Unsupervised Learning with Random Forest Predictors”[*http://labs.genetics.ucla.edu/horvath/RFclustering/RFclustering.htm*](http://labs.genetics.ucla.edu/horvath/RFclustering/RFclustering.htm)*, Journal of Computational and Graphical Statistics, Volume 15, Number 1, Pages 118–138DOI: 10.1198/106186006X94072*
24. Shi, T and Horvath S., “A Tutorial for RF Clustering”, fttp://labs.genetics.ucla.edu/horvath/RFclustering/RFclustering.htm
25. Smith, Lindsay, “A Tutorial on Principal Components”, 2002, www.sccg.sk/~h**a**l**a**dov**a**/princip**a**l\_comp**on**ents.pdf -
26. Venibles W and Ripley B, *Modern Applied Statistics With S-PLUS*, Springer 1999
27. Viaene, S. “Learning to Detect Fraud From Enriched Insurance Claims Data: Context, Theory and Application, PhD dissertation, Katholieke Universiteit Leuven”, 2002
1. The simulated data is based on a research database originally constructed to investigate claims that were suspected not to be legitimate, such as staged accidents and inflated damages. The term “fraudulent” is generally not used in referring to such claims as claims that meet the definition of criminal fraud are a very small percentage of claims. [↑](#footnote-ref-1)
2. A part of the research for this paper was to develop a dataset that can be made publically available for research on predictive modeling methods [↑](#footnote-ref-2)
3. See Wikipedia article on insurance fraud at www.wikipwdia.org and Coalition Against Insurance Fraud web site, www.insurancefraud.org [↑](#footnote-ref-3)
4. See Wikipedia article on insurance fraud. The article cites the Insurance Information Institute, however the Institute’s web site now refers to “questionable” claims rather than fraud and does not cite a particular percentage. [↑](#footnote-ref-4)
5. Thus both of these variables could be used to construct additional binary categorical variables indication whether a claim was legitimate or had some level of suspicion of being questionable [↑](#footnote-ref-5)
6. In the original research, claims specialist created a multi-category variable indicating 1) probable legitimate or evidence of of a questionable of the following kinds 2) excessive treatment, 3) opportunistic fraud, no injury, 4) opportunistic fraud, injury, 5) planned fraud [↑](#footnote-ref-6)
7. Lee Yang, in an email containing some R code provided a formula for the “scaled RIDIT” as follows: Scaled RIDIT = 2\*RIDIT – 1. This RIDIT under certain assumptions about P(X = Xi) is the RIDIT in Brockett at al (2003). Brocket and Levine in their 1977 paper show the two RIDITS are the same if the appropriate transformation is applied. [↑](#footnote-ref-7)
8. http://www.statisticalanalysisconsulting.com/using-ridits-to-assign-scores-to-categories-of-ordinal-scales/ [↑](#footnote-ref-8)
9. Some of the code used was supplied by Lee Yang [↑](#footnote-ref-9)
10. A factor analysis as well as a principal components procedure can be used to derive the PRIDITs [↑](#footnote-ref-10)
11. The analysis in the 2012 study used real data, i.e., the original 1993 AIB data [↑](#footnote-ref-11)
12. Hastie, Tibshirani, and Friedman describe how the estimate resulting from bagging is similar to a posterior Bayesian estimate. [↑](#footnote-ref-12)
13. The value shown in each node is the average value of the Suspicion variable which takes on the value 1 for legitimate and 2 for suspected questionable claim [↑](#footnote-ref-13)