STOCHASTIC SIGNALS AND SYSTEMS

REPORT

A Review of Statistical Classification Methods in Credit Risk Analysis

August 23, 2016

Nivetha Mahalakshmi Balasamy

Rutgers University

Contents

1 PROPOSAL

Financial organizations and banking industries often employ different methods to identify and classify customers into risk groups for credit scoring. Credit risk is the risk which stems from a borrower failing to make required payments or when a borrower intends to use future funds to make current debt payments. Decisions of rejecting or granting a loan and interest rates for loans are made by references to these risk groups. Therefore low risky borrowers are likely to get credit compared to borrowers on the high risk end. Effective estimation of credit risk has therefore become a critical component used by many financial organizations to measure risk.

The objective of this study is to review the many statistical tools that have been studied and implemented for intelligent credit risk analysis. A holistic view of the entire process includes data pre- processing, classification and providing classified output. Input data pre-processing is done to remove data redundancy and make the classification more efficient. The main challenge however is developing classification models to distinguish good borrowers from the bad ones while preserving accuracy. Various statistical methods and recently Artificial Intelligence methods like back propagation Neural Networks and Support Vector Machines have been used to make predictions. Many classifiers have been developed in the last few years and evidence suggests that Support Vector Machines[1] and Hidden Markov Models have the highest accuracy[2][3]. Although many classifiers have been presented in literature this study discusses the conventional techniques like Logistic regression, Linear Discriminant Analysis , Bayesian Behavioral models[4] along with superior classifiers like Support Vector Machines and Hidden Markov models for credit scoring. This project aims to organize the various results and present an integrated view of the trends in classification models for credit risk analysis along with identifying future trends where more research is likely to continue.

2 CREDIT SCORING

Credit scores are primarily used by Financial Institutions to classify applicants between good and bad risk classes. Several significant predictors in the dataset (variables or feature vectors) aid in making good prediction and with large amounts of predictors available for classifying, choosing the right predictors becomes essential . The advent of high processing computing and highly efficient Machine Learning algorithms have made building of highly accurate classification models viable. This has also made way for real-time analysis of data, drastically reducing the computational time to make a decision.

2.1 Statistical Learning

The first step in Statistical Learning is cleaning the data set in terms of missing data and other redundancies. This is followed by extraction of predictors which have significant impact on credit score of the customer. Domain knowledge of experts and feature selection algorithms play a crucial role in this step. D.J Hand et al.[5] present a Forward stepwise selection algorithm which sequentially and iteratively add the variables that result in greatest increase in predictive accuracy. The data is then divided into training and the test data. The Machine Learning algorithm is trained using training data and evaluated on testing data. The false interpretation of good customer as a bad customer or vice-verse, called the missclassification error rate, is the common parameter for for evaluating the performance of various Learning methods.

2.2 Dataset

In [3] [6] [7] the German and Australian credit dataset from the UIC learning repository is used since real time credit data sets are not very easily available. In [5] various methods to deal with data inconsistencies, such as substituting values for missing values or dropping incomplete vectors are presented. Also, In [2] methods have been considered to include previously rejected population and the process is commonly known as reject inference.

3 CLASSIFICATION OF CREDIT CUSTOMERS

3.1 Logistic and Kernel Logistic Regression:

Logistic Regression, an extension of Linear regression, models the probability of the response belonging to a particular category given input values of its predictors.

$$\text{P(predictor)= P(default=Yes|predictor)} \text{ [8]}$$

where *P(predictor)* denotes the probability of default corresponding to the particular predictor. In order to model the probabilities between 0 and 1 the *logistic function* is used

$$P(X) = \exp^{\beta_0+\beta_1 X} / 1 + \exp^{\beta_0+\beta_1 X} \text{ [8]}$$

To extend for p number of predictors

$$P(X) = \exp^{\beta_0+\beta_1 X+...+\beta_p X_p} / 1 + \exp^{\beta_0+\beta_1 X+...+\beta_p, X_p} \text{ [8]}$$

The model is fit to the dataset using maximum likelihood function and works well when the data is linearly separable. Non linear datasets can be classified with an extention of Logistic Regression as described in [7]. The Kernel method maps the predictors to a higher dimension feature space which appears as non linear in the input space. The *logit* model of Kernel Logistic Regression can be written in the form

$$g(x) = \beta^T \phi(x) \text{ [7]}$$

Various types of Kernel functions exist and [7] discusses the Radial Basis Function Kernel due to lesser computations and capability to handle relations when attributes are non linear.

$$K(x_i, x_j) = \exp(-||x_i - x_j||^2 / 2\sigma^2) \text{ [7]}$$

Where the $||x_i - x_j||$ represents the Eucledian space in the higher dimension and σ^2 the *tuning parameter*. In order to avoid overfitting of data as the input vectors are mapped to higher dimensions, the penalty constant $\lambda/2||\beta||^2$ is applied along with the training equation.

$$L(\beta)_r = L(\beta) + \lambda/2||\beta||^2 \text{ [7]}$$

The pair of tuning parameter values selected in [7] uses the Grid Search algorithm (5- fold Cross Validation) in order to find the best predictor.

3.2 Linear Discriminant Analysis:

Linear Discriminant Analysis (LDA) approximates the Bayes classifier. The Bayes theorem in view of LDA can be stated as follows

$$Pr(Y = k|X = x) = \pi_k f_k(x) / \sum_{l=1}^{K} \pi_l f_l(x) \text{ [8]}$$

The estimation of the prior probability and $f_k(x)$ (which is assumed to be Gaussian) is used to find the conditional probability of Y given X. The prior probability of each class can estimated by counting the frequency of a particular class label in the training set [2]. Also, the mean and variance of the Gaussian distribution of a particular class can be estimated. When the following estimations are applied to Bayes theorem, the equation is as follows

$$\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k) \text{ [8]}$$

This results in a linear boundary between the classes to be separated. The LDA has low error rate if the data set satisfies two of these model assumptions. The main advantage of this particular model is that it works well for classes with more two class variables.

3.3 Bayesian Networks:

[2] and [9] show that Bayesian Networks(BN) for a set of predictors is represented by B=<G,θ> where G is directed acyclic graph nodes and θ is the conditional probability distribution of parent nodes. Further, missing links between any two variables in the graph indicate insignificant relationships between the predictors. This method uses Bayes theorem to determine the posterior probability where good domain knowledge of prior probability and likelihood function decides the accuracy of the posterior probability [2]. The joint probability distribution in [2] is given as

$$P_B(x_1,...,x_n) = \prod_{j=1}^{n} P_B(x_j|\pi(x_j))$$

which makes the BN a suitable classifier. For say, four predictors

$$P(y|x_1,x_2,x_3,x_4) = P(y,x_1,x_2,x_3,x_4)/P(x_1,x_2,x_3,x_4) \text{ [2]}$$

In [2] a variant of BN classifier called the Markov Chain Monte Carlo (MCMC) Bayesian classifier was evaluated on the German data set. This resulted in model with high model interpretation and good predictive performance. The main advantage of this method as stated by Witold et al.[4] is its ability to accept incomplete data and provide reason for rejection of an applicant.

3.4 Classification using Hidden Markov Model:

Hidden Markov Model(HMM) is a Markov Process in which the states are hidden from the observer, but having correlation with many observable parameters. In credit scoring, classification of a customer is done based on the available credit scoring parameters.

Oguz et al.[3] have presented a discrete time HMM model for credit scoring where the continuous values of the dataset are sampled and assigned discrete categorical values. For this system, there are two states, (good or bad customer) and the transition is a Markov process. However, the states are hidden and we can only estimate the probabilities from the observable parameters in the dataset. The data is organized as a matrix , where each row vector corresponds to a customer and the columns

are the attributes or feature vectors which are the observable parameters. The key here is to make maximum use of the observable information to estimate the hidden states.

Therefore, there is a random process which is hidden and another observable process which provides the observation generated from the hidden process. The dataset matrix is the observation matrix B and the state transition matrix A is a 2X2 matrix (good and bad states) from which the initial state distributions π are obtained. The HMM is represented by λ = (A,B,π). In [3] eleven feature vectors are considered making the observation matrix 11X11. Srivastava et al. [10] have used an observation set of three values low, medium and high. For a state sequence of say, length four $X = (x_0, x_1, x_2, x_3)$ and observations $O = (O_1, O_2, O_3, O_4)$ the state sequence probability is calculated as

$$P(X,O) = \pi_{x_0} b_{x_0}(O_0) a_{x_0,x_1} b_{x_1}(O_1) a_{x_1,x_2} b_{x_2}(O_2) a_{x_2,x_3} b_{x_3}(O_3) \text{ [11]}$$

The next step is training. Training the HMM can have two possibilities. For fully observed data with no hidden values π, A and B can be calculated by counting frequencies. When there are hidden variables the initial HMM parameters should be estimated. Srivastava et al. [10] estimate the symbol generation probabilities based on the card holder's spending profile and the initial state transition probabilities which are assumed to be uniform. In [3] and [6] the estimation is by maximum likelihood.

After the initial HMM parameters are obtained the HMM parameters should be learnt. In [3] the two models are trained with data from good and bad customers using the Viterbi algorithm. In [6][10] the sequences are formed from the observation symbols from training dataset.Badreddine Benyacoub et al. [6] present the Baum-Welch algorithm (Combination of foward and backward algorithm) in more detail where the fundamental techniques used in their work was inspired by [3].

In [3][6][10]the observation sequence from training set and newly generated transaction observation sequence is given to the HMM to compute the likelihood or acceptance probability of belonging to either the good or bad credit score class. For example, In [10], for a sequence of symbols $O_1, O_2, ..., O_R$ the acceptance probability is computed from α_1 and α_2 as $\alpha_1 = P(O_1, O_1, ..., O_R|\lambda)$ and a new sequence α_2 formed from dropping O_1 and appending O_{R+1} $\alpha_2 = P(O_2, O_3, ..., O_{R+1}|\lambda)$

$$\text{Let } \Delta\alpha = \alpha_1 - \alpha_2$$

The customer is rejected if $\Delta\alpha/\alpha_1 \geq$Threshold

3.5 Support Vector Machine:

The Support Vector Machine is a popular classification technique which relies on finding a suitable hyperplane for dividing the p-dimensional space into two regions, with each region being assigned to a particular class. SVM is an extension of Maximal Margin Classifier which results in a margin that maximises the distance between the closest training observation and the hyperplane [8]. The Maximal Margin Classifier can fit only linear hyperplane. In order to accommodate the nonlinear boundaries, SVM uses Kernel function to map the input vectors to a high dimensional feature space. The hyperplane separating the two classes in high dimensional feature space results in a non-linear

boundary in the original space as q(x) = 0 , where q is quadratic [1]. SVM is also called a *soft margin* classifier as it introduces a penalty constant *C* to allow slight misclassifications which helps to avoid data overfitting [8]. The choice of Kernel depends on the type of dataset. The two popular kernel methods are Radial Basis and Polynomial Kernel [8]. SVM works the best when classes are well separated.

4 BEST MODEL

Since Statistical methods involve several trade-offs, the choice of best model depends on the questions which need to be addressed by the particular dataset. As stated in [5] the best model really depends on various factors like representation of data in datasets, selection of predictors and ability of these predictors to successfully separate data into two classes . For problems where Model interpretation weighs more than Prediction, Logistic Regression and Bayesian Networks would be the preferred model of choice. SVM is a better fit for classes that are distinctly separable and when the population of the predictors is Normally distributed, Linear Discriminant Analysis would be suitable. Hidden Markov Model allows a large number of predictors to be considered while providing a more deterministic approach. In the future as data about customer grows *Deep Learning* methods can be implemented to increase the over all performance.

REFERENCES

[1] U. Worrachartdatchai and P. Sooraksa, "Credit scoring using least squares support vector machine based on data of thai financial institutions," *Advanced Communication Technology, The 9th International Conference on (Volume:3),DOI: 10.1109/ICACT.2007.358779*, pp. 2067–2070, 2007.

[2] B. Baesens, "Developing intelligent systems for credit scoring using machine learning techniques," *Ph.D Thesis, Katholieke Universiteit leuven*, 2003.

[3] H. T. Oguz and F. S. Gurgen, "Credit risk analysis using hidden markov model," *Computer and Information Sciences, 2008. ISCIS '08. 23rd International Symposium, DOI: 10.1109/ISCIS.2008.4717932*, 2008.

[4] W. Abramowicz, M. Nowak, and J. Sztykiel, "Bayesian networks as a decision support tool in credit scoring domain," *The Poznan University of Economics*, 2003.

[5] D. Hand and W.E.Henley, "Statistical classification methods in consumer credit scoring," *J.R. Statsit. Soc.A 160,Part 3*, pp. 523–541, 1997.

[6] A. Z. Badreddine Benyacoub, Souad El Bernoussi, "Building classification models for customer credit scoring," *Logistics and Operations Management (GOL), International Conference DOI: 10.1109/GOL.2014.6887425*, pp. 107 – 111, 2014.

[7] S.P.Rahayu, J. M. Zain, A.Embong, and S. W. Purnamii, "Credit risk classification using kernel logistic regression with optimal parameter," *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pp. 602–605, 2010.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An introduction to statistical learning with applications in r,"

[9] T. Pavlenko and O. Chernyak, "Credit risk modeling using bayesian networks," *International Journal Of Intelligent Systems, VOL. 25*, pp. 326–344, 2010.

[10] S. S. Abhinav Srivastava, Amlan Kundu and A. K. Majumdar, "Credit card fraud detection using hidden markov model," *IEEE Transactions on Dependable and Secure Computing, VOL. 5, NO. 1,*, pp. 37–48, 2008.

[11] M. Stamp, "A revealing introduction to hidden markov models," *San Jose State University.*