Improved of K-Nearest Neighbor Techniques in Credit Scoring

Mr.Ram Babu.#1, Mr.A.Rama Satish#2

#1 Student, Dvr & Dr. Hs Mic College Of Technology, Kanchikacherla,Krishna(dt)

#2 Assoc. professor, Dvr & Dr. Hs Mic College Of Technology, Kanchikacherla,Krishna(dt)

#1 rambabu33216@gmail.com, #2 ramsatpm@gmail.com,

Abstract: Credit scoring has gained more and more attentions both in academic world and the business community today. Many modeling techniques have been developed to tackle the credit scoring tasks. Credit scoring models have been widely used by financial institutions to determine if loan customers belong to either a good applicant group or a bad applicant group. The advantages of using credit scoring models are K-Nearest Neighbor can be described as the benefit from reducing the cost of credit analysis, enabling faster credit decision, insuring credit collections. This model is compare with other models depends upon the each models credit scoring.

I.INTRODUCTION

The last two decades have seen a rapid growth in both the availability and the use of consumer credit. Until recently, the decision to grant credit was based on human judgment to assess the risk of default. The growth in the demand for credit, however, has led to a rise in the use of more formal and objective methods (generally known as credit scoring) to help credit providers decide whether to grant credit to an applicant . Credit scoring can be formally defined as a statistical (or quantitative) method that is used to predict the probability that a loan applicant or existing borrower will default or become delinquent. This helps to determine whether credit should be granted to a borrower.

Credit scoring can also be defined as a systematic method for evaluating credit risk that provides a consistent analysis of the factors that have been determined to cause or affect the level of risk. The objective of credit scoring is to help credit providers quantify and manage the financial risk involved in providing credit so that they can make better lending decisions quickly and more objectively. Credit scoring is a system creditors use to assign credit applicants to either a "good credit" one that is likely to repay financial obligation or a "bad credit" one who has high possibility of defaulting on financial obligation.

Credit scoring has many benefits that accrue not only to the lenders but also to the borrowers. Credit scoring also helps to increase the speed and consistency of the loan application process and allows the automation of the lending process. As such, it greatly reduces the need for human intervention on credit evaluation and the cost of delivering credit. With the help of the credit scores, financial institutions are able to quantify the risks associated with granting credit to a particular applicant in a shorter time. Further, credit scores can help financial institutions determine the interest rate that they should charge their consumers and to price portfolios. Higher-risk consumers are charged a higher interest rate and vice versa. Based on the consumer's credit scores, the financial institutions are also able to determine the credit limits to be set for the consumers. These help financial institutions to manage their accounts more effectively and profitably. Because of advances in technology, more intelligent credit scoring models are being developed. Consequently, credit card issuers are able to make use of the information generated from the models to formulate better collection strategies and hence use their resources more effectively.

 ISSN: 2320-7884 (Online)

In this paper, we look at the application of K-Nearest Neighbor, a standard technique in pattern recognition and non parametric statistics to the credit scoring problem. The K-NN method involves in estimating good or bad risk probabilities for an applicant to be classified by the proportions good and bad among the k most similar points in the training samples . The similarity of points accessed by the suitable distance metric

II RELATED WORK

*Model tree (M5) :*The M5 model tree is a numeric prediction algorithm. M5 algorithm generates a conventional decision-tree structure with linear regression models at the leaves, which would give a smoother score distribution instead of discrete class labels. The basic tree is generated by a recursive partition method similar to C4.5. Then, a regression function is built for each node of the constructed tree using the standard regression algorithm. The regression functions are simplified by minimizing an estimate of the expected error, which is multiplied by a factor (n+v)/(n-v). Here n is the number of training cases that reach that node, v is the number of terms (including the constant and the independent variables) in the linear regression function at that node. Terms in the linear model are dropped one by one, greedily, so long as doing so decreases Eerror. Thus, due to the factor (n+v)/(n-v), the linear model may be simplified to minimize the Eerror. Finally, once the final simplified linear model is placed for each node, the tree is pruned back from the leaves. If the Eerror of a node is smaller than the Eerror of the subtree below, the subtree is replaced by this single node. A parameter of M5 algorithm is the pruning factor F, which decides the extent of the pruning. Trees with different size are generated by varying this parameter.

Multi-layer perceptron neural network with back-propagation (MLP): MLP consists of the input, hidden and output layers of interconnected nodes. The nodes in the network are all sigmoid. Through multiple passes of training with the examples, the weights of the nodes are modified, based on the error rates of the resulting outputs. Only one hidden layer is used in the network in the following experiments. The number of hidden nodes (H) is the only parameter to be determined. Network models have other parameters: learning rate (set to be 0.3) and momentum (set to be 0.2). If the network diverges from the answer, the network will be automatically reset with a lower learning rate and be trained again.

Logistic Regression (LR): Logistic regression can predict the probability (P) that an example X belongs to one of two predefined classes. Suppose example X=(x1, x2,…, xk), as in linear regression, logistic regression gives each xi a coefficient wi which measures the contribution of each xi to variations in P. First, a logistic transformation of P is defined as logit(P)= ln(P/(1 – P)) where P can only range from 0 to 1, while logit (P) ranges from -∞ to ∞. Logit (P) is then matched by a linear function of the feature variables:

$$\text{logit}(P)=\ln(P/(1-P))= w0 + w1x1 + w2x2 + w3x3 +\ldots.+ wkxk$$

Parameters wi are usually estimated by the maximum likelihood method.

III K-Nearest Neighbor Method

The k-NN algorithm in this study follows the method described in Aha et al.3 In this method, a set of training examples is saved. A Euclidean distance metric is used to measure the similarity between each training example and a new example. In this simple k-NN algorithm, features will not be weighted and they are treated equivalently. The algorithm finds the k examples nearest to the new example. The new example is assigned to the class to which the majority of the k neighbour examples belong. Neighbours will be weighed by the inverse of their distance when voting. The number of neighbours (parameter k) is the important factor. A suitable k is to be empirically determined for the data set used in order to smooth the influence of the noise in the data set.

The Selection of appropriate distance measure is an important part of K-NN method. A standard distance measure is Euclidean distance is given by:

$$D(X,Y)= \{(X-Y)^T (x-Y)\}^{1/2}$$

ISSN: 2320-7884 (Online)

Where X and Y are points in feature space.

The data used in our analysis consist of a sample room the full population of applicants for credit. All the applicants in the sample, including applicants who would normally have been rejected, were given credit and observed over a period of a year. A standard procedure was used to define creditworthiness. Applicants who defaulted for three consecutive months were classified as bad and the remaining applicants in our data set were classified as good. This allows us to treat creditworthiness as a two-state problem and to use techniques for dealing with binary response data. The predictor variables used for constructing credit scoring models were extracted room the application form and other sources, such as a historical database. These variables, which we call characteristics, are usually nominal or ordinal in nature. . The best bound for the bad risk rate among those accepted is given by

$$\text{Best BR} = \begin{cases} 1-p/a & a>p, \\ 0 & a=p. \end{cases}$$

The worst bound for the bad risk rate among the acceptances is given by

$$\text{WORST BR} = \begin{cases} (1-p)/a & a>1-p \\ 1 & a\leq 1-p \end{cases}$$

Corresponding bounds on the best and worst good risk rates for the rejects and error rates can also be found:

$$\text{Best GR} = \begin{cases} 0 & a>p \\ (p-a)/(1-a) & a\leq p \end{cases}$$

$$\text{WORST BR} = \begin{cases} 1 & a>1-p \\ p/(1-a) & a\leq 1-p \end{cases}$$

IV PERFORMANCE

Automatic updating of classification rule. One potential attraction of the k-NN method is that dynamic updating of the design set is easy. As (recently) accepted customers reach the point where their performance is assessed as good or bad, they can be added to the design set to replace the oldest surviving observation.

Reasons for refusal of credit. The k-NN method with adjusted Euclidean metrics could provide a reason for refusal of credit by exploiting the information about class separation in the data provided by the regression weights. To do this we could calculate the (standardized) distance between a new applicant and the mean for the design set bad applicants for each characteristic in the model. This gives an approximation to the contribution of each characteristic to the decision. Characteristics with low values of this distance are the characteristics which identify the applicant with previous bad applicants and, thus, ensure that the applicant is rejected. Of course, with any multivariate classification technique (including linear or logistic regression) it is not possible o reduce a decision explicitly to the value of a single variable.

Red lining of applicants. A charge sometimes levelled at credit grantors is that their methods of screening applicants red line people (refuse credit on the basis of one characteristic regardless of all other attributes). We believe that the e-NN method is less susceptible to this criticism than linear or logistic regression because our proposed distance metric takes into account random variation as well as the distance along the equiprobability contours (estimated room the regression). However, as the value of D increases, the k-NN decision becomes increasingly dependent on characteristics with large weights wt room the regression. One way to reduce this problem would be to put a limit on the ratio of the regression weights wf. This constraint would be likely to reduce the classification accuracy of the regression score-card by more than that of the k-NN classifier.

The incremental case analysis of k-NN proves the expected phenomena: the error rate is decreased with the increased sizes of the train sample. To show the further decreasing of error rates, twice of the train

ISSN: 2320-7884 (Online)

sample are reselected from the available cases and used to train models (see points of 200%). The trends of lines show that as the parameter k is increased, the error is decreased at first. However, when k is larger than 7, increasing of k cannot decrease error rates but rather increases error rates slightly. This result conforms to the expected shape of learning curve for incremental complexity analysis.

V APPLICATIONS

In the early years, financial institutions used credit scoring mainly to make credit decisions for loan applications. Over the past 25 years, however, the application of credit scoring has grown from making credit decisions to making decisions related to housing, insurance, basic utility services, and even employment. However, not all these applications are equally widely used.

The most common use of credit scores is in making credit decisions for loan applications. In addition to decisions on personal loan applications, financial institutions now make use of credit scores to help set credit limits, manage existing accounts, and forecast the profitability of consumers and customers.

Credit scoring models have also been used in the insurance industry (e.g., for mortgage and automobile insurance) to decide on the applications of new insurance policies and the renewals of existing polices. The premise is that there is a direct relationship between financial stability and risk. It has been argued that there is a strong relationship between credit rating and loss ratios in both automobile and mortgage insurance.

In addition to the above, other credit scoring applications have also been reported. For example, landlords can make use of credit scores to determine whether potential tenants are likely to pay their rent on time. There is substantial use of credit scoring in the mortgage industry too. Also, some utility suppliers in the United States have used credit scores to determine whether to provide their services to consumers. Finally, some employers make use of credit history and credit scores to decide whether to hire a potential employee, especially for posts where employees need to handle huge sums of money

VI CONCLUSION

Credit scoring models are known as statistical models which have been widely used to predict the default risk of individuals or companies. In this paper, K-nearest neighbor is proposed for credit scoring. K-NN is a nonparametric classifier based on learning by similarity. A training data set is collected, for this training data set, a distance function is introduced between the explanatory variable of observations. For each new observation this method explores the pattern space for the K nearest neighbors that are closest to the new observation in term of distance between the explanatory variables. The new observation is assigned to the class which its most KNN belong to that class.

VII REFERENCES

1. Chen, M. C., and Huang, S. H., 2003, "Credit scoring and rejected instances reassigning through evolutionary computation techniques." Expert Systems with Applications, 24(4), 433–441.
2. Apilado, V P., Waner, D. C. and Dauten, I. J. (1974) Evaluative techniques in consumer finance—experimental results and. policy implications. J. Finan. Quant. Anal., Mar., 275-283.
3. Boyle, M, Crook, J. N., Hamilton, R. and Thomas, L C. (1992) Methods for credit scaing applied to slow payers.InProc. Conf Credit Scoring and Credit Control (eds L. C. Thomas, I. N. Crook and D. B. Edelman), pp. 75-90. Oxford: Clarendon.
4. Otgler, Y. E. (1970) A credit scoring model for commercial loans. J. Money Credit Bank., Nov., 435-445.
5. Paredes, R., and Vidal, E., 2000, "A class-dependent weighted dissimilarity measure for nearest neighbor classification problems." Pattern Recognition Letters 21(12), 1027-1036.
6. Hand, D. J., and Vinciotti, V., 2003, "Choosing k for two-class nearest neighbor classifiers with unbalanced classes." Pattern Recognition Letters 24(9-10), 1555-1562.
7. Islam, M. J., Wu, Q. M. J., Ahmadi, M., and Sid-Ahmed, M. A., 2007, "Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers"International Conference on Convergence Information Technology. IEEE Computer Society.
8. Marinakis, Y., Marinaki, M., Doumpos, M., and Matsatsinis, N., 2008, "Constantin Zopounidis, Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment." Journal of Global Optimization 42(2), 279-293.
9. Li, F. C., 2009b, "The Hybrid Credit Scoring Model based on KNN Classifier"Sixth International Conference on Fuzzy Systems and Knowledge Discovery. IEEE Computer Society.