Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Credit scoring analysis using weighted k nearest neighbor

To cite this article: M A Mukid *et al* 2018 *J. Phys.: Conf. Ser.* **1025** 012114

Credit scoring analysis using weighted k nearest neighbor

M A Mukid, T Widiharih, A Rusgiyono, A Prahutama

Department of Statistics, Diponegoro University, Semarang, Indonesia

E-mail: mamukid@live.undip.ac.id

Abstract. Credit scoring is a quatitative method to evaluate the credit risk of loan applications. Both statistical methods and artificial intelligence are often used by credit analysts to help them decide whether the applicants are worthy of credit. These methods aim to predict future behavior in terms of credit risk based on past experience of customers with similar characteristics. This paper reviews the weighted k nearest neighbor (WKNN) method for credit assessment by considering the use of some kernels. We use credit data from a private bank in Indonesia. The result shows that the Gaussian kernel and rectangular kernel have a better performance based on the value of percentage corrected classified whose value is 82.4% respectively.

Keywords: credit scoring, weighted k nearest neighbor, nonparametric classification

1. Introduction

Credit is an important catalyst for economic growth and is a core activity of banks around the world. According to The Hong Kong Institute of Bankers [1], the success or failure of a bank and a financial industry generally depends on the system used to manage the credit and how well the credit risk is handled. The availability of credit allows households to perform better consumption and allow companies to make investments that can not be done with own funds. But with moral hazard and adverse selection issues, banks play an important role in allocating capital and monitoring to ensure that public funds are channeled to activities that provide optimal benefits [2]. One way for capital allocation to reach the target is to make predictions about the ability to pay future customers.

Credit risk level assessment methods have played an important role in the practice of contemporary banking risk management. They contribute to the key to a loan approval process that accurately and efficiently quantifies the credit risk level of a prospective borrower. These credit assessment methods aim to predict future behavior in terms of credit risk based on past experience of customers with similar characteristics. The level of a borrower's credit risk is attributed to the chance that it will default on an approved loan at a predetermined time. The main task of a credit scoring method is to provide a separation between those who fail and those who do not fail in terms of credit payments. The separating ability is a key indicator of a method's success [3].

Several banchmarking studies have also been conducted to compare empirically the performance of such techniques in estimating credit scores, as did by Baesens et al. [4]. They compared seventeen models using eight real sets of data and it is known that more complex techniques tend to produce better performance based on Area Under Curve (AUC) criteria. De-La-Vega et al. [5] compared linear discriminant analysis models, quadratic discriminant analysis, logistic regression, multilayer perceptron, SVM, tree classification and combined methods to data from a microfinance institution and concluded that the use of multilayer perceptron is better than models other. The large number of bancmarking studies on credit rating models often leads to sometimes conflicting conclusions. Yobas, Crook, and Ross [6] found that analysis discriminant linear were better than artificial neural network while Desai, Crook, and Overstreet [7] reported that artificial neural network were significantly better than analysis discriminant linear. Until now, it is not clear what literature states that there is an appropriate model for credit assessment [8].

Research on credit risk level assessment models has also evolved taking into account the conditions of available sample data. Bucker, Kampen and Kramer [9] build a credit scoring model with regard to missing data conditions. Their research yields a conclusion that one of them is that the alleged parameter values differ significantly both statistically and economically when compared to cases where consumers whose credit rejected are ignored. Niklis, Doumpos and Zopounidis [10] built a credit scoring model based on information from market and accounting information. Brown and Mues build credit scoring models with due regard to conditions where the number of "good credit" and "bad credit" cases is not balanced [8]. Usually many cases of "bad credit" are much less than in many cases of "good credit".

One of the most commonly used methods for credit scoring is k nearest neighbor (KNN). This method belongs to the category of nonparametric classification method. It is known that the non-parametric classifier usually suffer from the existing outliers, especially in the situation of small training sample size [11]. There are many credit scoring researchers that has used KNN to asses the risk associated with the lending to an organization or an individual [12]. Henley and Hand used KNN classifier for assessing consumer credit scoring by proposing an adjusted version of the Euclidean distance metric [13]. KNN is also used as a comparative method of new methods proposed by researchers in credit scoring ([14],[8]). Up to now there is no researcher who applied weighted k nearest neighbor (WKNN) methods for credit scoring. The WKNN method assigns different weights to each of the nearest neighbes k. The closer neighbors will get a greater weight than the more distant neighbors. The weighting process is done by using the kernel function. We are interested in reviewing the performance of WKNN for credit scoring analysis compared with conventional KNN.

The rest of this paper is organized as follows. In Section 2, we give a brief overview of many algorithms, including KNN and WKNN. The data used in this paper describe in Section 3. A study of credit scoring analysis from a financial institution and its results are presented in Section 4. Finally, conclusions are offered in Section 5.

2. Overview of k Nearest Neighbor and Its Modified Version

This study aims to compare the performance of k nearest neighbor and its modified version within a credit scoring context. A brief explanation of each of the methods applied in this study is presented below.

2.1 K Nearest Neighbor (KNN)

KNN is a non parametric lazy learning algorithm. It is a non parametric technique, it means that it does not make any assumptions on the underlying data distribution. This is very useful, as in the real world, most of the practical data does not obey the typical theoretical assumptions made. Non parametric algorithms like KNN come to the rescue here. Most of the lazy algorithms – especially KNN – makes decision based on the entire training data set.

KNN is one of the simplest of classification algorithms that often used as a benchmark for more complex classifiers. Fix and Hodges introduced a method for pattern classification that has since become known the k-nearest neighbor rule [15]. It is a nonparametric method, which means that it does not make any assumptions about the probability distribution of the input. The k-nearest neighbor classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. The main idea of k-NN algorithm is that whenever there is a new point to predict, its k nearest neighbors are chosen from the training data. Then, the prediction of the new point can be the average of the values of its k nearest neighbors [16].

2.2 Weighted k Nearest Neighbor (WKNN)

WKNN is one of the election rules where different members of the nearest neighboring group are weighted by the distance function between the training data and the test data. WKNN uses the same principle of KNN as finding the closest distance between the data to be tested with a number of k nearest neighbor in the training data. WKNN gives the heaviest weight on the nearest neighbor and the smallest on the farthest neighbor according to distance function [17]. The working principle of WKNN follows the working principle KNN i.e. looking for test data with the closest distance to train data according to the selected neighbor. WKNN change the distance value on the nearest ladder to a value between 0 and 1. The closest distance will be assigned a value 1. Conversely, the farthest distance will be assigned a value of 0.

WKNN is an extension version of KNN where the distances of the nearest neighbors can be taken into account. WKNN gives the closest neighbors weights greater than other farther neighbors. The process of weightinh is done through two stages, namely first counting distance and the second transforming the distance into a weight by using a kernel function. Hechenbichler and Schliep has developed this method by transforming the similarity measure as the weights [18]. Below is the WKNN algorithm:

Let L = $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n_L\}$be a learning set of observation $\mathbf{x}_i$ with the class membership y_i and let $\mathbf{x}$ be a new observation, whose class label y has to predicted.

1. Find the k + 1 nearest neighbors to x based on $d(\mathbf{x},\mathbf{x}_i)$
2. Standardize the k smallest distance via
$$D_{(i)} = D(\mathbf{x},\mathbf{x}_{(i)}) = \frac{d(\mathbf{x},\mathbf{x}_{(i)})}{d(\mathbf{x},\mathbf{x}_{(k+1)})}$$
3. Transform $D_{(i)}$ with any kernel function K(.) into weights $w_{(i)} = K(D_{(i)})$
4. Predict the class membership y of observation $\mathbf{x}$ by choosing the class, which shows a weighted majority of the k nearest neighbors $\hat{y} = \max_r \left(\sum_{i=1}^{k} w_{(i)} I(y_{(i)} = r) \right)$

Hechenbichler dan Schliep [18] noted some kernel functions that can be used in WKNN i.e.

1. Kernel Epanechnikov: $\frac{3}{4}(1 - D^2) \cdot I(|D| \leq 1)$
2. Kernel quartic atau biweight: $\frac{15}{16}(1 - D^2)^2 \cdot I(|D| \leq 1)$
3. Kernel triweight: $\frac{35}{32}(1 - D^2)^3 \cdot I(|D| \leq 1)$

   where $I = \begin{cases} 1, & |D| \leq 1 \\ 0, & |D| > 1 \end{cases}$
4. Kernel gauss: $\frac{1}{\sqrt{2\pi}}\exp(-\frac{D^2}{2})$
5. Kernel inverse: $\frac{1}{|D|}$

3. Material and Method

This paper uses data that comes from a bank in Indonesia consisting 948 clients of which 184 clients are categorized as bad customer. The bank defines that a bad customer is a someone who had missed three consecutive months of payments. The data consist 8 continuous explanatory variables including age, working experince, total income, other loan, net income, interaction to bank, savings, and debt ratio. The coding, descriptive of each variables and unit of measure were shown in Table 1.

Table 1. Variables in credit scoring model

Variable	Definition
Age	Age of the applicant in years
Working Experience	Working experience of the applicant in years
Total Income	Monthly income in Rupiahs
Other Loan	Other loan amount in Rupiahs
Net Income	Monthly net income of the applicants in Rupiahs
Interaction To Bank	Long interaction to bank in years
Saving	Amount of the applicants savings in Rupiahs
Debt Ratio	% per month

Table 1 explains that some variables have unit of measurement different from others. Therefore to calculate the distance between two objects, the variables used in the analysis are standardized first using formula $z_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$ where x_{ij} is jth observation of ith variable, $\bar{x}_i$ is mean of ith variable, and s_i is standart deviation of ith variable. For this study, we divide the data into training sample consisting of 80% and testing sample. We use KNN and WKNN for classifying the class of the debtors. Next, calculate the euclide distance between two objects using standardized data. After determining the parameter k and running the algorithm of each method, the accuracy was calculated using sensitivity, specificity, and accuration (ACC). The formulas of the three measures of accuracy can be seen in [19].

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Accuration} = \frac{TP+TN}{TP+FP+TN+FN}$$

where *TP* (true positive) is the number of good borrower that were classified as good; *FN* (false negative) is the number of good borrower that were mistakenly classified as bad; *TN* (true negative) is the number of bad borrower that were classified as bad; *FN* (false negative) is the number of observed bad customer that were were mistakenly classified as good.

4. Results and Discussions

Tables 2 and 3 are statistical overview of good and bad customers respectively based on 8 explanatory variables. The tables explain that the average of each variable on good customers show to be better than bad customers. For instance, the average of total monthly income of good customers reaches Rp 10278655.530 which is greater than the total revenue of bad customers whose value only reaches Rp 8855448.065. Furthermore, the mean of nominal savings from good customers reach Rp 15065382.662 that is greater than bad customers. Similar condition occur to other variables that showed a good customer has a positif individual performance. Tables 2 and 3 also describe that there are a different variation in the predictor variables involved in the analysis. Some of the variables in good customers have a greater variability than in bad ones but for other variables the variability is lower than that of bad ones. For example the working experience variability of good customers is greater than the age variability of bad customers but the age variability of good customers is lower than that of bad ones.

Table 2. Statistical summary of a good customer

Variables	Mean	Standart Deviation	Minimum	Maximum
Age	37.726	7.606	23.000	64.000
Working Experience	7.665	6.781	1.000	39.000
Total Income	10278655.530	12089028.731	2253967.200	144000000.000
Other Loan	358474.500	1194111.250	0.000	17392868.000
Net Income	5295639.944	5980694.323	0.000	63786168.000
Interaction to Bank	3.881	2.884	0.000	19.000
Saving	15065382.662	124231063.340	18599.400	2880181083.600
Debt Ratio	30.608	11.054	7.485	77.069

Table 3. Statistical summary of a bad customer

Variables	Mean	Standart Deviation	Minimum	Maximum
Age	37.038	8.331	22.000	57.000
Working Experience	6.190	6.608	1.000	34.000
Total Income	8855448.065	9809392.959	2275227.000	88471800.000
Other Loan	237887.793	1020949.380	0.000	9512016.000
Net Income	4668156.853	4966941.559	0.000	49151000.000
Interaction to Bank	2.549	2.244	0.000	10.000
Saving	7318661.557	21114280.029	40266.000	176765400.000
Debt Ratio	32.496	14.979	8.282	130.724

Table 4. Accuration of WKNN using some kernels

Kernel	Accuration Measure	Parameter K						
		1	3	5	7	9	11	13
Rectangular	Sensitivity	0.8618	0.8947	0.9539	0.9605	0.9737	**0.9934**	1.0000
	Spesifisity	0.3056	0.2500	0.1667	0.1667	0.0833	**0.1111**	0.0556
	Accuration	0.7553	0.7713	0.8032	0.8085	0.8032	**0.8245**	0.8191
Triangular	Sensitivity	0.8618	0.8750	0.9013	0.9408	0.9539	0.9671	0.9671
	Spesifisity	0.3056	0.3333	0.2222	0.1944	0.1667	0.1389	0.1389
	Accuration	0.7553	0.7713	0.7713	0.7979	0.8032	0.8085	0.8085
Epanechnikov	Sensitivity	0.8618	0.8816	0.9079	0.9474	0.9539	0.9671	0.9671
	Spesifisity	0.3056	0.3333	0.1944	0.1944	0.1944	0.1667	0.1667
	Accuration	0.7553	0.7766	0.7713	0.8032	0.8085	0.8138	0.8138
Triweight	Sensitivity	0.8618	0.8816	0.8750	0.8947	0.9079	0.9145	0.9342
	Spesifisity	0.3056	0.3056	0.3056	0.2778	0.2222	0.1944	0.1944
	Accuration	0.7553	0.7713	0.7660	0.7766	0.7766	0.7766	0.7926
Gaussian	Sensitivity	0.8618	0.8947	0.9539	0.9605	0.9671	**0.9934**	1.0000
	Spesifisity	0.3056	0.2500	0.1667	0.1667	0.0833	**0.1111**	0.0556
	Accuration	0.7553	0.7713	0.8032	0.8085	0.7979	**0.8245**	0.8191
Inversion	Sensitivity	0.8618	0.8882	0.9539	0.9539	0.9605	0.9803	1.0000
	Spesifisity	0.3056	0.2500	0.1667	0.1667	0.1111	0.1111	0.0556
	Accuration	0.7553	0.7660	0.8032	0.8032	0.7979	0.8138	0.8191

In this paper we examine the performance of several kernels including rectangular, triangular, Gauss, epanechnikov, triweight, and inversion for classifying credit status of customers. The main task of a kernel function is to transform the distance of a new object to its nearest neighbor to a weight. The neighbor is closer than others will gain a greater weight.

Table 4 explain the accuration measure of WKNN in different both kernel and parameter k. In order to find the best parameter value k we tried some values of k and evaluated its classification capability based on sensitivity, specificity and accuration. In this paper we consider only odd value of k due to reason of classifying. We choose the combination between kernel and k as a main ingredients of WKNN if its accuration is maximum. From table 4, we know that the kernel rectangular and Gauss at k = 11 achieve the highest PCC values (82.4%) among the other kernels. At k =11, both kernel rectangular and Gauss have the sensitivity value 99.34% and the specificity value 11.11%. For k lower or greater than 11, all accuration measure smaller than of k = 11. It mean that overall 82.4% of debtors are properly classified. Then 99.3% of good debtors by WKNN are classified as good and 11.1% of bad debtors by WKNN are classified as bad. After finding the right value for k and choosing the kernel used, the next step is to predict the new loan applicant whether he or she is eligible to receive the credit or not. The first step is to calculate the distance between the new loan applicant and the debtor that the bank has. The Euclidian distance is calculated based on the variables as shown in Table 1. Furthemore, by using rectangular or Gauss kernel, we transform the distance into a weight. Then we identify the 11 borrowers closest to the new loan applicants. If the weight of a good debtor is greater than the weight of the bad debtor then the new loan applicant is classified as a good consumer and is eligible for approval of his credit proposal.

5. Conclusion

Credit scoring model is the most successful example of statistical model applied in financial institution. The objective of quantitative credit scoring models is to assign credit applicants to one of two group: a “good” group that is likely to repay their financial or a “bad” group that should be denied credit because a high likelihood of defaulting on their financial obligation. In this paper, we explore weighted k nearest neighbor method for classifying credit applicants into good or bad. The main raw material of WKNN is a kernel function that change the distance between two object into a weight. We investigate the performance of some kernel function including rectangular, triangular, Gauss, epanechnikov, triweight, and inversion for classifying credit status of customers. We found that kernel rectangular and Gauss are the relevant kernel used for credit scoring analysis. In our case, their optimal performance reach at k = 11.

Acknowledgment

This research was supported by the research and technology and higer eduction ministry of Indonesia based on contract number 007/SP2H/LT/DRPM/IV/2017.

References

[1] The Hong Kong Institut of Bankers. 2012. Credit risk management. Wiley: Hong Kong.

[2] Utari, G. A. D., Arimurti, T., & Kurniati, I. N. 2012. Pertumbuhan kredit optimal. *Buletin Ekonomi Moneter dan Perbankan*. 10, 1-34.

[3] Nicolic, N., Zarkic-Joksimovic, N., Stojanovski, D., &Joksimovic, I., 2013. The application of brute force logistic regression to corporate credit scoring models: Evidence from Seerbian financial statements. *Expert Sytems with Applications*, 40, 5932-5944.

[4] Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. 2003. Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, 54(6), 627–635.

[5] De La Vega, M. D. C., Oliver, A. B., Mejias, R. P., & Rubio J. L., 2013. Improving the management of microfinance institutions by using penilaian kredit models based on statistical leraning techniques. *Expert System with Applications*, 40, 6910-6917.

[6] Yobas, M. B., Crook, J. N., & Ross, P., 2000. Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, 11(2), 111–125.

[7] Desai, V.S., Crook, J.N., & Overstreet, G.A., 1996. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research* 95, 24–37.

[8] Brown, I., & Mues, C., 2012. An experimental comparison of classification algorithm for imbalanced credit scoring data set. *Expert Sytems with Applications*, 39, 3446-3453.

[9] Bucker M., Kampen M. V., & Kramer W., 2013. Reject Inference in Consumer Credit Scoring with Nonigorable Missing Data. *Journal of Banking & Finance*, 37, 1040-1045.

[10] Niklis D., Doumpos M., & Zopounidis C., 2014. Combining Market and Accounting-Based Models for Credit Scoring Using a Classification Scheme Based on Suport Vector Machines. *Applied Mathematics and Computation*, 234, 69-81.

[11] Fukunaga, K., (1990). *Introduction to Statistical Pattern Recognition*, second ed. Academic Press.

[12] Paleologo, G., Elisseeff, A., & Antonini, G., 2010. Subagging for credit scoring models. *European Journal of Operational Research*, 201, 490-499.

[13] Henley, W. E., & Hand D. J., 1996. A k nearest-neighbour classifier assessing consumer credit risk. *The Statistician*, 45(1), 77-95.

[14] Lessmanna, S., Baesens, B., Seowd, H. V., & Thomas, L. C., 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247, 124-136.

[15] Fix, E., & Hodges, J. L., 1951. Discriminatory analysis nonparametric discrimination: Consistency properties. *Technical Report TR4*, USAF School of Aviation Medicine, Randolph Field, TX.

[16] Zhang Y., & Wang J., 2016. K-nearest neighbors and a kernel density estimator for GEFCom2014 probabilistic wind power forecasting. *International Journal of Forecasting*, 32, 1074-1080.

[17] Gou, J., Yi, Z., Du, L., & Xiong, T., 2012. A local mean-based k-nearest cetroid neighbor classifier. *The Computer Journal*, 55(9).

[18] Hechenbichler dan Schliep. 2004. *Weighted K Nearest Neighbor Techniques and Ordinal Classification*. Sonderforschungsbereich 386, Paper 399.

[19] Louzada, F., Ferreira-Silva, P. H., & Diniz, C. A. R, 2012. On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data. *Expert Sytems with Applications*, 39, 8071-8078.