Using Memory-Based Reasoning For Predicting Default Rates On Consumer Loans

Jozef Zurada, (E-mail: jmzura01@louisville.edu), University of Louisville
Robert M. Barker, (E-mail: rmbark01@louisville.edu), University of Louisville

ABSTRACT

In recent years, financial institutions have struggled with high default rates for consumer lending. An ability to reliably predict the probability of consumer loan defaults would have a significant impact of the profitability of that lending for these institutions. In response to this need, the financial institutions have employed loan analysis techniques such as logistic regression, discriminant analysis, and various machine learning techniques to improve the accuracy of detecting loan defaults. The objective of these techniques is to more precisely identify creditworthy applicants who are granted credit, thereby increasing profits, from non-creditworthy applicants who would be then denied credit, thus decreasing losses. The objective of this article is to employ an emergent data analysis technique, memory-based or case-based reasoning method, to this problem to test its accuracy in discriminating between good and bad loans. This paper examines historical data from consumer loans issued by a financial institution to individuals that the financial institution considered to be qualified customers. The data set consists of the financial attributes of each customer and includes a mixture of loans that the customers paid off or defaulted upon. The paper then compares the performance of this technique to other data mining techniques proposed in earlier works and analyzes the risk of default inherent in each loan for each technique.

INTRODUCTION

The consumer credit market in the United States is estimated by the US Federal Reserve to be a $7 trillion market (Musto and Souleles, 2006). In the last several years, the financial services industry has experienced a rapid growth with significant increases in mortgages, auto-financing, debts to retailers, credit card debts, and home equity loans to name a few. The Federal Reserve estimates that there is currently more than $5.4 trillion in mortgage debt, $800 billion in installment debt, and $700 billion in revolving debt (Musto and Souleles, 2006). With this growth, however, have been mounting losses for delinquent and defaulted loans. For example, in 1991, $1 billion of Chemical Bank's $6.7 billion in real estate loans were delinquent and the bank held $544 million in foreclosed property. Manufacturers Hanover's $3.5 billion commercial property portfolio was burdened with $385 million in non-performing loans (Rosenberg and Gleit, 1994). The FHA was found to have mortgage default rates in 8 of 22 cities studied of more than 12.7%, with one lender having a 34.1% default rate in Buffalo, N.Y. (National Mortgage News, 2002). In 1994 in the UK about 12% of retail expenditure was made using credit cards, amounting to a total of about 36 billion British pounds (Hand and Henley, 1997). This dynamic is also highly relevant for the former East-block Central and East European countries, now members of the European Union. For example, in 2004 alone, Czek and Slovak banks recorded 33.8% and 36.7% increases in their retail loans, respectively (Vojtek and Kocenda, 2006). In response, many financial services institutions have placed increasing importance on the development of new credit scoring models in addition to attempts to increase the robustness of traditional statistical techniques to support their credit decisions. The ultimate objective of these models is to increase accuracy in loan-granting decisions, so that more creditworthy applicants are granted credit, thereby increasing profits, and non-creditworthy applicants are denied credit, thus decreasing losses. A slight improvement in accuracy translates into significant future savings, which can have an important impact on profitability.

Unfortunately, credit-risk evaluation decisions are inherently complex and unstructured. This is due to the nonlinear relationships between independent variables that interact with each other and the various forms of risks involved. The most harmful risk to the party approving credit is the nonpayment of obligations when they come due. Simultaneously, the payoff associated with a correct credit-risk decision is high. Due to the difficulty of credit risk assessment, the financial institution that provided data for this paper experienced a default rate of about twenty percent (20%), even though that financial institution used a complex credit scoring model to attempt to identify bad loans. Some defaults are caused by unforeseen factors (i.e. stability of marriage, health status and/or job stability) that may be difficult to reflect in the financial attributes collected by the bank about the consumer. However, some bad loans could be avoided by using more discriminating machine-based credit risk assessment techniques. Possible bad loans could be flagged for a more thorough analysis by a human loan officer, who could then scrutinize the potential loan further for its credit worthiness. As a result, any improvement in making a reliable discrimination between those who are likely to repay the loan and those who are not would be highly desired.

This paper examines and compares the effectiveness of memory-based reasoning model (*k*-nearest neighbor) to predict whether a consumer defaulted upon or paid off a loan. Memory-based reasoning is a data mining technique that has garnered considerable attention as a means to understand dynamic behavior in complex data (Im and Park, 2007). From the original, highly unbalanced data set containing 5960 loan applicants, which was heavily dominated by good loans (4771 and 1189 cases constituted good and bad loans, respectively), we randomly selected 1189 good loans and matched them with the same number of bad loans. We repeated this procedure 10 times to get more mileage out of the data and make the results more generalizable. As a result, each of the 10 data sets used in this study contained an equal mix of 1189 of randomly selected good loans and all available 1189 bad loans. We used the memory-based reasoning technique, ran computer simulation for 8 different sizes *k* of the neighborhood for each of the 10 data sets, and recorded the results for 3 probability cutoffs: 0.3, 0.5., and 0.7. The classification results were averaged across 10 test data sets and standard deviation was computed to test the stability/dispersion of the obtained results. The memory-based reasoning method turned out efficient at identifying both good loans and bad loans in the test sets given that the financial institution that provided the data considered all of the loans contained in the data set to be good loans warranting an extension of credit. The paper assesses and analyzes the probability of default on a single loan and compares the obtained results to the results published in the earlier study by (J. Zurada and M. Zurada, 2002).

The paper is organized as follows. Section 2 covers memory-based reasoning technique fundamentals. Section 3 reviews the current literature. Section 4 describes the data sample used in this study. Section 5 presents the experiments and simulation results. Finally, section 6 concludes the paper and gives some recommendations for future work.

MEMORY- OR CASE-BASED REASONING METHOD

Memory-based reasoning is a type of case-based reasoning. Broadly construed, it is the process of solving new problems based on the solutions of similar past cases. The memory-based reasoning method requires no model to be fitted, or function to be estimated. Instead it requires all cases with their known solutions to be maintained in memory, and when a prediction is required, the method recalls items from memory and predicts the value of the dependent variable. In solving a new case, the memory-based reasoning approach retrieves a case it deems sufficiently similar and uses that case as a basis for solving the new case.

Memory- or case-based reasoning employs a *k*-nearest neighbor algorithm to classify cases. The *k*-nearest neighbor algorithm takes a data set of existing cases and a new case to be classified, where each existing case in the data set is composed of a set of variables and the new case has one value for each variable. The normalized Euclidean distance or Hamming distance between each existing case and the new case (to be classified) is computed. The *k* existing cases that have the smallest distances to the new case are the *k*-nearest neighbors to that case. Based on the target values of the *k*-nearest neighbors, each of the *k*-nearest neighbors votes on the target value for a new case. The votes are the posterior probabilities for the class dependent variable (LOAN_SATUS).

More formally, a high-level summary algorithm that computes the distance between each new case $z = (\mathbf{x}', y')$ and all the training patterns $(\mathbf{x}, y) \in D$ to calculate its nearest-neighbor list, D_z, can be outlined as follows (Tan *et al*., 2006).

1. Let *k* be the number of nearest neighbors and *D* be the set of training patterns.
2. for each new case $z = (\mathbf{x}', y')$ do
3. Compute $d(\mathbf{x}', \mathbf{x})$, the distance between *z* and every pattern, $(\mathbf{x}, y) \in D$.
4. Select $D_z \subseteq D$, the set of *k* closest training patterns to *z*.
5. $y' = \underset{v}{\text{argmax}} \sum_{(x_i, y_i) \in D_z} I(v = y_i)$
6. end for.

Once the nearest neighbors list is obtained, the new case is classified based on the majority class of its nearest neighbors:

Majority voting: $y' = \underset{v}{\text{argmax}} \sum_{(x_i, y_i) \in D_z} I(v = y_i)$

where *v* is a class label, y_i is the class label for one of the nearest neighbors, and $I(\cdot)$ is an indicator function that returns the value 1 if its argument is true and 0 otherwise.

In the majority voting approach, every neighbor has the same impact on the classification. This makes the algorithm more sensitive to the choice of *k*. To reduce the influence of *k*, one can weight the impact of each nearest neighbor $\mathbf{x}_i$ according to its distance: $w_i = 1/d(\mathbf{x}', \mathbf{x}_i)^2$. As a result, training patterns that are located far away from *z* will have a smaller influence on the classification compared to those that are located closer to *z*. Using the distance-weighted voting scheme, the class label of the new case can be determined as follows:

Distance-Weighted Voting: $y' = \underset{v}{\text{argmax}} \sum_{(x_i, y_i) \in D_z} w_i \times I(v = y_i)$

Table 1 explains the above algorithm in more practical terms. It illustrates the majority voting approach for a binary dependent variable, LOAN_STATUS, by the nearest neighbors when different values of *k* are specified. (The nearest neighbors are not distance-weighted.) Panel A presents an example where the existing cases 6, 15, 35, 68, 123, 178, 190, 205, 255, and 269 are the 10 closest cases to the new case which we want to classify. Cases 35 and 123 have the shortest and the longest distances to the new case, respectively. The *k*-nearest neighbors are first *k* cases that have the closest distances to the new case.

Panel B illustrates how the posterior probability for a new case is determined. If the value of *k*=3, the target values of the first two nearest neighbors (35, 190, and 6) are used. The target values for these 3 neighbors are Repaid (0), Defaulted (1), and Repaid (0). Therefore, the posterior probability for the new case to have the target value Repaid (0) and Defaulted (1) is 2/3 (66.7%) and 1/3 (33.3%), respectively. If *k*=5, the target values for the 5 nearest neighbors (35, 190, 6, 205, and 269) are used. The target values for these 5 neighbors are Repaid (0), Defaulted (1), Repaid (0), Repaid (0), and Repaid (0), respectively. Therefore, the posterior probability for the new case to have the target value Repaid (0) and Defaulted (1) is 4/5 (80%) and 1/5 (20%), respectively. Similarly, if *k*=10, the target values for the 10 nearest neighbors (35, 190, 6, 205, 269, 255, 178, 15, 68, and 123) are used. The target values for these 10 neighbors are Repaid (0), Defaulted (1), Repaid (0), Repaid (0), Repaid (0), Repaid (0), Defaulted (1), Defaulted (1), Repaid (0), and Repaid (0), respectively. Therefore, the posterior probability for the new case to have the target value Repaid (0) and Defaulted (1) is 7/10 (70%) and 3/10 (30%), respectively.

Table 1
Illustration Of The *K*-Nearest Neighbor Classification Algorithm For The Dichotomous Target Variable LOAN_STATUS Taking Two Values Repaid And Defaulted.

Panel A. Description Of 10 Nearest Neighbors (Sorted By Case Number).

Case id	Target attribute: LOAN_STATUS (Repaid or Defaulted)	Case ranking based on the distance to the instance: (1 – closest, 10 – farthest)
6	Repaid	3
15	Defaulted	8
35	Repaid	1
68	Repaid	9
123	Repaid	10
178	Defaulted	7
190	Defaulted	2
205	Repaid	4
255	Repaid	6
269	Repaid	5

Panel B. Example Of Posterior Probability Determination.

k	Case id of nearest neighbor(s)	Target value of nearest neighbor(s)	Posterior probability **P** of the instance
3	35, 190, 6	Repaid, Defaulted, Repaid	**P** (Repaid)=66.7% **P**(Defaulted)=33.3%
5	25, 190, 6, 205, 269	Repaid, Defaulted, Repaid, Repaid, Repaid	**P**(Repaid)=80% **P**(Defaulted =20%
10	25, 190, 6, 205, 269, 255, 178, 15, 68, 123	Repaid, Defaulted, Repaid, Repaid, Repaid, Repaid, Defaulted, Defaulted, Repaid, Repaid	**P**(Repaid)=70% **P**(Defaulted)=30%

One can easily notice that there are two critical choices in the nearest neighbor-method, namely, the distance function and the cardinality *k* of the neighborhood. We performed 8 experiments for different values of *k* and used the normalized Euclidean distance for numeric variables and the Hamming distance for categorical variables to calculate the similarity between cases in 12-dimensional space (12 independent variables). Normalization was required to ensure that features with larger values do not overweight features with lower values. Furthermore, to minimize the influence of *k*, we used the voting approach with weighted-distance in computer simulation. For more details on the memory-based reasoning method, the reader is encouraged to refer to (Mitchell, 1997; Han and Kamber, 2001; Giudici, 2003; SAS Enterprise Miner: http://www.sas.com; Tan *et al*., 2006; and Im and Park, 2007).

PRIOR RESEARCH

Financial institutions are very reliant on the use of computer-based reasoning for extension of credit to consumers. Use of such techniques as a decision aid to loan officers helps to increase the speed and accuracy of the lending process. During recent years many articles have been published describing the effectiveness of machine learning tools such as neural networks, decision trees, fuzzy systems, genetic algorithms, rough sets, and regression analysis in predicting customer loan default rates. For example, two papers by Rosenberg and Gleidt (1994) and Hand and Henley (1997) are chiefly concerned with whether to extend credit to an applicant. The aim is to anticipate and reduce defaults and serious delinquencies. Other credit risk management concerns discussed in the two papers are the maintenance of existing credit lines (ie. should the credit limit be raised?) and determining the best action to be taken on delinquent accounts.

In more recent articles focusing on loan risk assessment, Barney *et al*. (1999) compared the performance of neural networks and regression analyses in identifying the farmers who had defaulted on their Home Administration Loans and those farmers who paid off the loans as scheduled. Using an unbalanced data, Barney found that neural networks outperform logistic regression in correctly classifying farmers into those who made timely payments and

those who did not. Jagielska *et al*. (1999) investigated credit risk classification abilities of neural networks, fuzzy logic, genetic algorithms, decision trees, and rough sets and concluded that the genetic/fuzzy approach compared more favorably with the neuro/fuzzy and rough set approaches. In two of his papers, Piramuthu (1999) analyzed the beneficial aspects of using both neural networks and nuero-fuzzy systems for credit-risk evaluation decisions. Piramuthu used three real-world applications data that involved credit-risk evaluation in various forms: credit approval, loan default, and bank failure prediction. Neural networks performed significantly better than neuro-fuzzy systems in terms of classification accuracy, on both training as well as testing data. However, the neural network cannot explain the rationale behind its credit granting/denial decision, unlike the neuro-fuzzy systems that explain decisions using simple if-then rules.

West (2000) investigated the credit scoring accuracy of five neural network architectures and compared them to traditional statistical methods. The neural architectures and traditional models included multilayer perceptron, mixture-of-experts, radial basis function, learning vector quantization, and fuzzy adaptive resonance; and discriminant analysis, logistic regression, *k* nearest neighbor, kernel density estimation, and decision trees, respectively. Using two real world data sets and testing the models using 10-fold cross validation, the author found that among neural architectures the mixture-of-experts and radial basis function did best, whereas among the traditional methods regression analysis was the most accurate.

Thomas (2000) surveyed the techniques for forecasting financial risk of lending to consumers, Yang *et al*. (2001) examined the application of neural networks to an early warning system for loan risk assessment, and Zurada *et al*. (2002) reported some preliminary results comparing the performance of data mining techniques in predicting the credit worthiness of customers. Onorato and Altman (2005) estimated loan default risk as a "mortality rate", using an actuarial-based approach to estimate loan transition risk. Feldman and Gross (2005) applied decision trees for detecting mortgage default rates. Also, Musto and Soules (2006) studied consumer credit risk using a covariance technique to compare credit default risk with aggregate consumer default rates in a portfolio context, finding significant heterogeneity of covariance risk across consumers.

Memory-based reasoning technique has been used for assessing consumer credit risk. For example, Hanley and Hand (1996), Hand and Henley (1997), and Hand and Vinciotti (2003) state that non-parametric nature of the method enables modeling of irregularities in the risk function over the feature space. Also, it is a fairly intuitive procedure and as such it could be easily explained to business managers who would need to approve its implementation. They emphasize that the choice of *k* (the size of the neighborhood) and the distance metric are challenging problems. The memory-based reasoning technique was also successfully applied in other business problem domains. Radding (1997) suggests the use of this technique for real-time fraud detection in long-distance telephone services, using variables such as frequency of calls, time of day, duration, and geography. Pequeno (1997) also proposes having several options, including neural networks and memory-based reasoning, available for detecting telecommunications fraud. Both latter authors maintain, however, that no single technique solves every problem and that using multiple techniques is the rule. In other words, different algorithms will have different levels of success, depending on the particular problem and the data. Memory-based reasoning, therefore, offers another potential avenue to help identify behavioral patterns in complex and dynamic data.

DATA SET USED IN THE STUDY

To build classification models, researchers have used data derived from various loan-granting contexts. West (2000) used a data set that contained 24 independent variables and 1000 observations collected on 1000 credit applicants to an important bank in southern Germany. The data set consisted of 700 records of customers who paid the loan off and 300 records of customers who defaulted. Fahrmeir and Hamerle (1994) and Giudici (2003) also analyzed the above German scoring data set with 1000 observations, but they reduced it to 21 variables. Desai *et al.* (1996) used 18 independent features and 3 data sets of about 900 cases describing the ordinary customers of three credit unions. The Australian scoring data (Quinlan, 1987) were similar but more balanced with 307 and 383 observations of each outcome. All data sets used in modeling such problems have an inherent bias as they contain only those individuals actually given a loan. There are others that did not get a loan and we do not know whether or not they would have been at risk. In other words, all data sets used in analysis contained observations about

customers that banks deemed to be qualified (creditworthy) individuals. Although these considerations do not affect the validity of the analysis, we should keep them in mind.

In our study we use qualitative variables that are similar to the ones used in the previous studies. We use a sample data provided by a money lending institution. The data set contains financial information about 5960 consumers allocated among 13 variables. The financial institution extended loans to all of the applicants in the data set since all of them seemed to be qualified customers. Those applicants who were denied a loan during the application process were not included in the data sample which we investigated. Out of the 5960 applicants, 4771 had paid off their loans and 1189 defaulted on the loans, resulting in a default rate of approximately 20%. We used tree imputation method to replace missing values. Out of these 13 variables, there were 12 independent variables (loan and consumer characteristics) and one dependent/target variable (loan default or loan repaid) that we are to predict. Using this data set, we build several memory-based reasoning models to predict whether a future applicant will default upon a loan or pay it off. The data set contains the following 12 independent variables: (1) Amount of the current loan request, (2) Amount due on existing mortgage, (3) Value of current property, (4) Debt-to-income ratio, (5) Years on current job, (6) Number of major derogatory credit reports, (7) Number of credit lines, (8) Number of delinquent credit lines, (9) Number of recent credit inquiries, (10) Age (in months) of oldest trade line, (11) Reason for loan (debt consolidation or home improvement), and (12) Applicants' job category. The binary dependent (target) variable Loan_Status takes values of 1 (client defaulted on loan or seriously delinquent) or 0 (loan repaid).

EXPERIMENTS AND RESULTS

Since the *k*-nearest neighbor method computes distances (similarities) between samples, we normalized each of the 12 independent variables to *z*-scores (with the arithmetic mean = 0 and variance = 1). Thus, the features with higher values do not overweight the features with lower values.

We performed a stratified sampling to create 10 balanced data sets by randomly selecting 1189 loans from the 4771 good loans and matching them with always the same 1189 bad loans. This sampling process produced 10 data sets, each consisting of 2378 cases divided evenly among good and bad loans. In each of the 10 data sets, we allocated 70% and 30% of the cases to the training set and test data set. As a result, the training and test sets contained 1663 and 715 observations, respectively. We built the models on the training set. The test set is used to calculate the performance of the models in the real world environment. To obtain an unbiased estimate of the future performance of the models and generalize the results, we ran computer simulation for 10 different generations of the balanced data sets, each containing 2378 cases and averaged the classification results.

On each of the 10 data sets, we then performed 8 experiments for different values of *k*. To calculate the similarity between cases in 12-dimensional space (12 independent variables), we used the normalized Euclidean distance for numeric variables and the Hamming distance for categorical variables. Also, to minimize the influence of *k* (the size of the neighborhood), we used the voting approach with weighted-distance in computer simulation.

The aggregated results from computer simulation are presented in Tables 1 and 2 and Figures 1 through 6. From computer simulation we recorded the averaged overall, repaid, and defaulted correct classification accuracy rates all their standard deviations [all in % format] for the three cutoff probabilities, i.e., 0.3, 0.5, and 0.7. The choice of the best model may depend on what cutoff probability a financial institution chooses to use. Because the target event was detecting loan defaults, the 0.3 cutoff implies that the cost of making an error of granting a loan when it should not be granted is 3.3 times higher than the cost of denying a loan when it should be granted. This cutoff may be applicable to situations in which banks do not secure smaller loans, i.e., do not hold any collateral. The 0.5 cutoff means that the cost of making an error of granting a loan when it should be denied is equal to the cost of denying a loan when it should be granted. Consequently, the 0.7 cutoff implies that the cost of making an error of granting a loan when it should be denied is 3.3 times smaller than the cost of denying a loan when it should be granted. The latter two cutoffs may typically be used when a financial institution secures larger loans by holding collateral such as title on a car or house purchased by the customer.

It appears that for the 0.3 cutoff, detecting bad loans is paramount. Medium size neighborhoods k=16 and k=20 work best in this instance and they detect 89.3% and 89.7%, respectively, of bad loans correctly. Also, for these two neighborhoods, the standard deviations for the averaged correct classification rates for loan defaults across 10 different generations of test sets are stable and equal to 1.9% and 1.5 %, respectively. Overall and repaid classification accuracy rates are the highest for smaller sizes of neighborhoods (k=5, k=8, and k=12).

For the 0.5 cutoff, the overall, repaid, and defaulted classification accuracy rates are the highest for k=5 (77.3%), k=5 (80.8), and k=8 (78.3%), respectively. These results are not significantly different from the results reported for the same cutoff probability in the study by J. Zurada and M. Zurada (2002). Averaged standard deviations are between 1% and 2%.

Smaller (k=5) and medium size neighborhoods (k=16) work also best for the 0.7 probability cutoff yielding 73.7%, 94.1%, and 55.7% for the overall, repaid, and defaulted classification accuracy rates, respectively.

CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

The memory-based reasoning method can be used dynamically by adding loan applicants when their class (repaid, defaulted) becomes known and deleting old applicants to overcome problems with changes in population over time. The results obtained in this study seem to be consistent with a simulation study performed by Enas and Choi (1986) and observations made by Hand and Vinciotti (2003) The former suggest that the k has to be much smaller than the smallest class ($k=n^{2/8}$ or $n^{3/8}$ is reasonable). In our case, the size n of the smallest class (loan defaults) is 715. Thus, the range for k is [5,12].

Further research should focus on refining the training and testing of the memory-based reasoning models using various balanced and unbalanced data sets to improve the classification performance. Once the models are fine tuned, one could attempt to adjust the classification results by the prior probabilities in the original data set (20% loan defaults), and establish the most profitable lending policy from a credit risk perspective based on the projected profits on good loans, average losses on bad loans, and the fixed and variable costs of lending operations.

REFERENCES

1. Barney, D.K., Graves, O.F., and Johnson, J.D., The Farmers Home Administration and Farm Debt Failure Prediction, *Journal of Accounting and Public Policy*, Vol. 18, pp. 99-139, 1999.
2. Desai, V.S., Crook, J.N., and Overstreet, G.A.Jr, A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment, *European Journal of Operation Research*, Vol. 95, pp. 24-37, 1996.
3. Enas, G. and Choi, S., Choice of the Smoothing Parameter and Efficiency of k-nearest Neighbor Classification, *Computers and Mathematics with Applications*, Vol. 12, 235-244, 1986.
4. Fahrmeir, L. and Hamerle, A., *Multivariate Statistical Modeling Based on Generalized Linear Programs*, Springer-Verlag, Berlin, 1994.
5. Feldman, D. and Gross, S., Mortgage Default: Classification Tree Analysis, *Journal of Real Estate Finance and Economics*, Vol. 30, 369-396, 2005.
6. FHA Pressed to Step Up Enforcement Against High-Default Lenders, *National Mortgage News*, Vol. 26, Issue 35, 3, May 7, 2002.
7. Giudici, P., *Applied Data Mining: Statistical Methods for Business and Industry*, John Wiley & Sons Ltd., Chichester, West Sussex, England, 2003.
8. Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, San Francisco, CA, 2001.
9. Hand, D. and Henley, W., Statistical Classification Method in Consumer Scoring: A Review, *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, Vol. 160, 523-541, 1997.
10. Hand, W. and Vinciotti, V., Choosing k for Two-Class Nearest Neighbor Classifiers with Unbalanced Classes, *Pattern Recognition Letters*, Vol. 24, 1555-1562, 2003.

11. Henley, W. and Hand, D., A *k*-nearest Neighbor Classifier for Assessing Consumer Credit Risk, *Statistician*, Vol. 45, 77-95, 1996.
12. Im, K. H. and Park, S. C., Case-based Reasoning and Neural Network Based Expect System for Personalization, *Expert Systems with Applications*, Vol. 30, Issue 1, 77-85, 2007.
13. Jagielska, I., Matthews, C., and Whitfort, T., An Investigation into the Application of Neural Networks, Fuzzy Logic, Genetic Algorithms, and Rough Sets to Automated Knowledge Acquisition for Classification Problems, *Neurocomputing*, Vol. 24, pp. 37-54, 1999.
14. Mitchell, T.M., *Machine Learning*, WCB/McGraw-Hill, Boston, Massachusetts, 1997.
15. Musto, D. and Soules, N. A Portfolio View of Consumer Credit, *Journal of Monetary Economics*, Vol.53, Issue 1, 59-84, 2006.
16. Onorato, M. and Altman, E., An Integrated Pricing Model for Defaultable Loans and Bonds, *European Journal of Operations Management*, Vol. 163, Issue 1, 65-82, 2005.
17. Pequeno, K. 1997. Real-time fraud detection: Telecom's next big step. *Telecommunications* 31 (5): 59-60.
18. Piramuthu, S., Financial Credit-Risk Evaluation with Neural and Neurofuzzy Systems, *European Journal of Operational Research*, Vol. 112, 310-321, 1999.
19. Piramuthu, S., Feature Selection for Financial Credit-Risk Evaluation Decisions, *INFORMS Journal on Computing*, Vol. 11, No. 3, pp. 258-266, 1999.
20. Quinlan, J.R., Simplifying Decision Trees, *International Journal of Man-Machine Studies*, Vol. 27, pp. 221-234, 1987.
21. Radding, A. 1997. Unpacking the mystery of the black box. *Software Magazine*, December 1997: S8-S9.
22. Rosenberg, E., and Gleidt, A., Quantitative Methods in Credit Management: A Survey, *Operations Research*, Vol. 42, 589-613, 1994.
23. Tan, P-N., Steinbach, M., and Kumar, V., *Introduction to Data Mining*, Addison-Wesley, 2006.
24. Thomas, L.C., A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers, *International Journal of Forecasting*, Vol. 16, pp. 149-172, 2000.
25. Vojtek, M. and Kocenda, E, Credit Scoring Methods, *Czech Journal of Economics and Finance*, Vol. 56, Issue 3-4, 152-167, 2006.
26. West, D., Neural Network Credit Scoring Models, *Computers & Operations Research*, Vol. 27, pp. 1131-1152, 2000.
27. Yang, B., Li, L.X., Ji, H., and Xu, J., An Early Warning System for Loan Risk Assessment Using Artificial Neural Networks, *Knowledge-Based Systems*, Vol. 14, pp. 303-306, 2001.
28. Zurada, J. and Zurada, M., How Secure are Good Loans: Validating Loan-Granting Decisions and Predicting Default Rates on Consumer Loans, *The Review of Business Information Systems*, 6(3), pp. 65-83, 2002.

Table 1
Averaged Correct Classification Accuracy Rates (Counts And Percentages)
For Different Sizes *K* Of Neighborhood.

k	5	8	12	16	20	25	30	40
				cutoff=0.3				
Overall	540	530	517	511	506	510	507	504
	75.5	74.0	72.2	71.3	70.6	71.2	70.8	70.4
Repaid	230	221	202	191	185	195	186	183
	64.2	61.7	56.4	53.4	51.7	54.4	51.9	51.2
Defaulted	310	309	316	320	321	315	321	321
	86.7	86.2	88.3	89.3	89.7	87.9	89.7	89.6
				cutoff=0.5				
Overall	554	547	540	538	534	532	532	524
	77.3	76.4	75.4	75.2	74.5	74.3	74.3	73.1
Repaid	289	267	269	270	271	279	275	275
	80.8	74.4	75.1	75.3	75.6	78.0	76.8	76.8
Defaulted	264	280	271	270	263	253	257	249
	73.9	78.3	75.8	75.3	73.5	70.6	71.7	69.5
				cutoff=0.7				
Overall	528	524	508	495	507	486	490	479
	73.7	73.2	70.9	69.2	70.8	67.9	68.4	66.8
Repaid	328	328	334	337	331	337	335	338
	91.7	91.7	93.4	94.1	92.5	94.1	93.6	94.4
Defaulted	200	196	173	158	176	149	155	140
	55.8	54.7	48.4	44.2	49.1	41.6	43.2	39.2

Table 2
Standard Deviations In [%] Of Averaged Correct Classification Rates (For Counts And Percentages) For Different Sizes *K* Of Neighborhood.

k	5	8	12	16	20	25	30	40
				cuttoff=0.3				
Overall	8.7	10.2	11.4	6.8	12.6	11.0	10.1	8.6
	1.2	1.4	1.6	0.9	1.8	1.5	1.4	1.2
Repaid	5.7	6.9	8.9	6.7	11.4	12.1	12.1	10.4
	1.6	1.9	2.5	1.9	3.2	3.4	3.4	2.9
Defaulted	5.9	6.3	7.3	6.6	5.5	6.5	4.7	5.4
	1.7	1.8	2.1	1.9	1.5	1.8	1.3	1.5
				cuttoff=0.5				
Overall	7.9	7.2	9.1	10.3	10.2	12.5	13.4	14.4
	1.1	1.0	1.3	1.4	1.4	1.8	1.9	2.0
Repaid	8.3	6.2	7.9	7.3	8.2	6.7	7.3	9.0
	2.3	1.7	2.2	2.0	2.3	1.9	2.0	2.5
Defaulted	9.3	6.8	10.5	8.2	10.0	11.6	10.2	12.5
	2.6	1.9	2.9	2.3	2.8	3.2	2.8	3.5
				cuttoff=0.7				
Overall	6.2	7.7	12.6	8.3	10.3	7.2	10.5	13.2
	0.9	1.1	1.8	1.2	1.4	1.0	1.5	1.8
Repaid	7.3	7.1	5.3	4.3	3.7	3.4	4.8	4.0
	2.0	2.0	1.5	1.2	1.0	0.9	1.3	1.1
Defaulted	8.0	8.3	11.3	7.6	10.5	6.9	10.9	12.3
	2.2	2.3	3.1	2.1	2.9	1.9	3.0	3.4

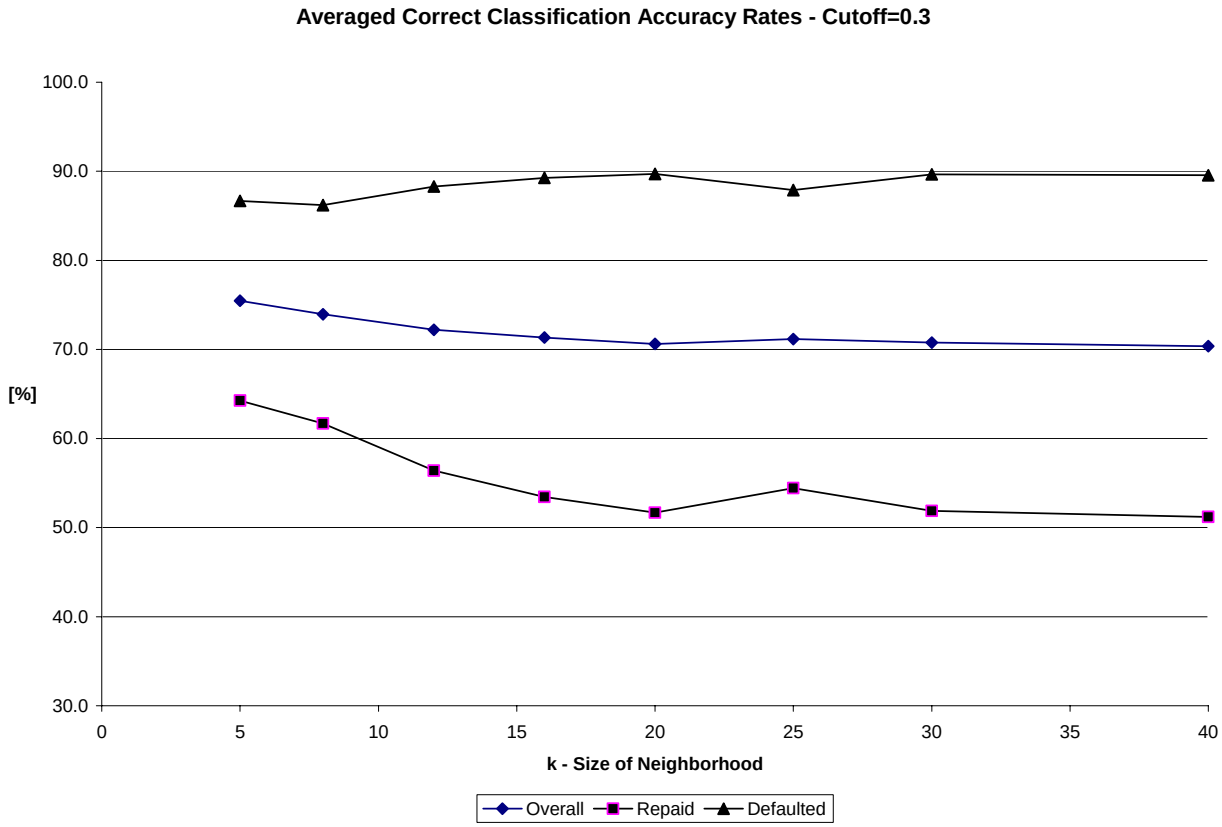

Figure 1
Averaged Overall, Repaid, And Defaulted Correct Classification Accuracy Rates [In %]
For The Cutoff Probability = 0.3

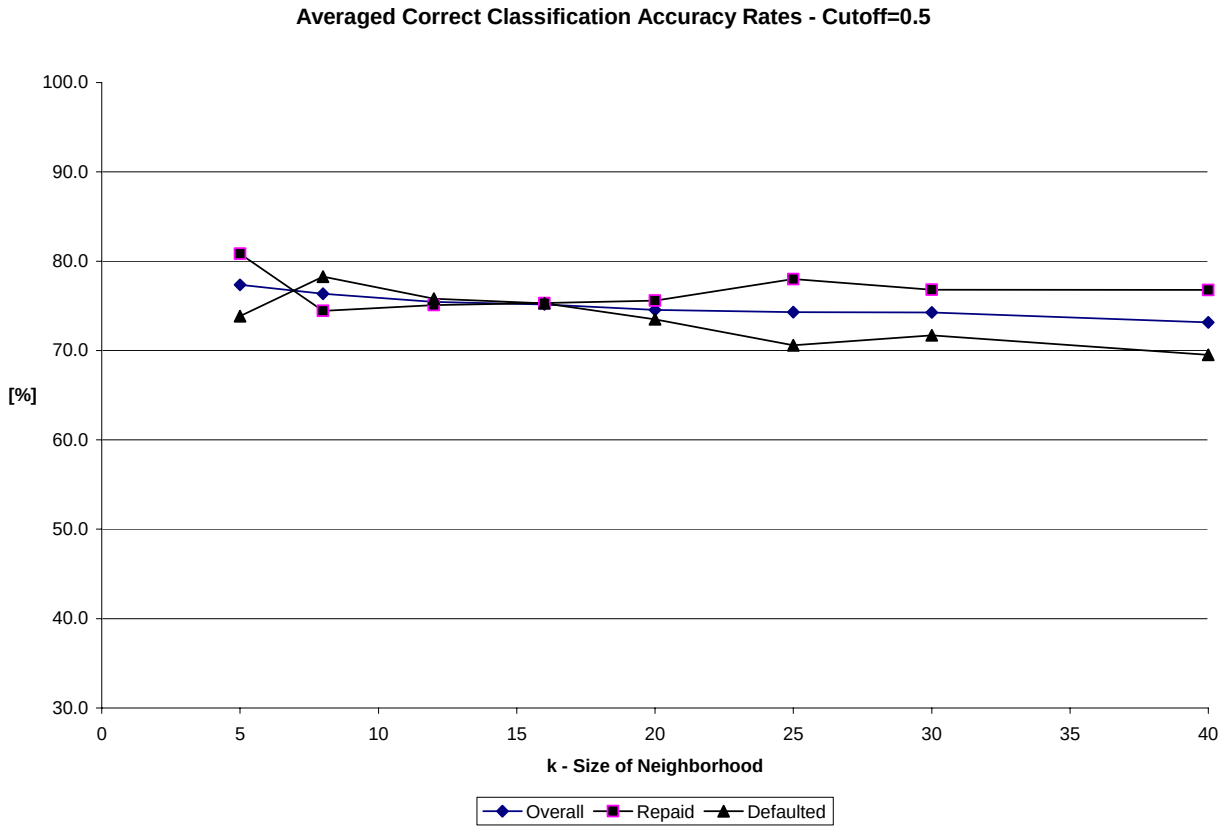

Figure 2
Averaged Overall, Repaid, And Defaulted Correct Classification Accuracy Rates [In %]
For The Cutoff Probability = 0.5

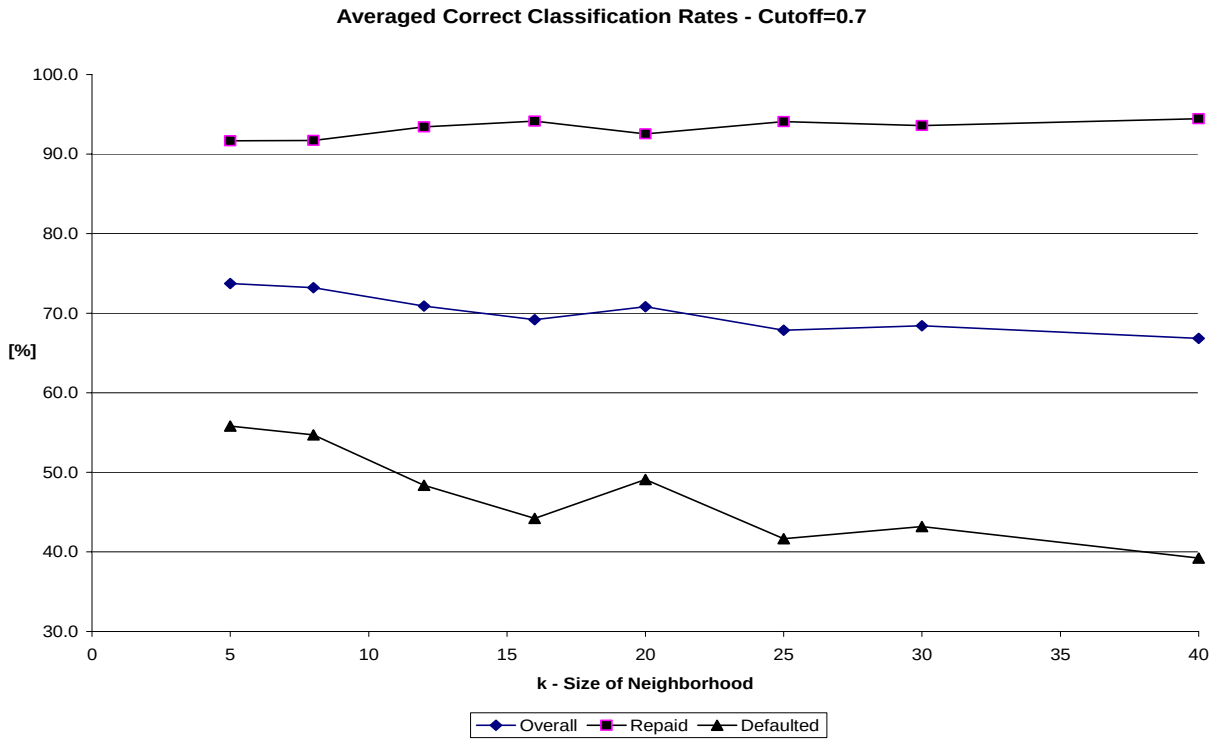

Figure 3
Averaged Overall, Repaid, And Defaulted Correct Classification Accuracy Rates [In %]
For The Cutoff Probability = 0.7

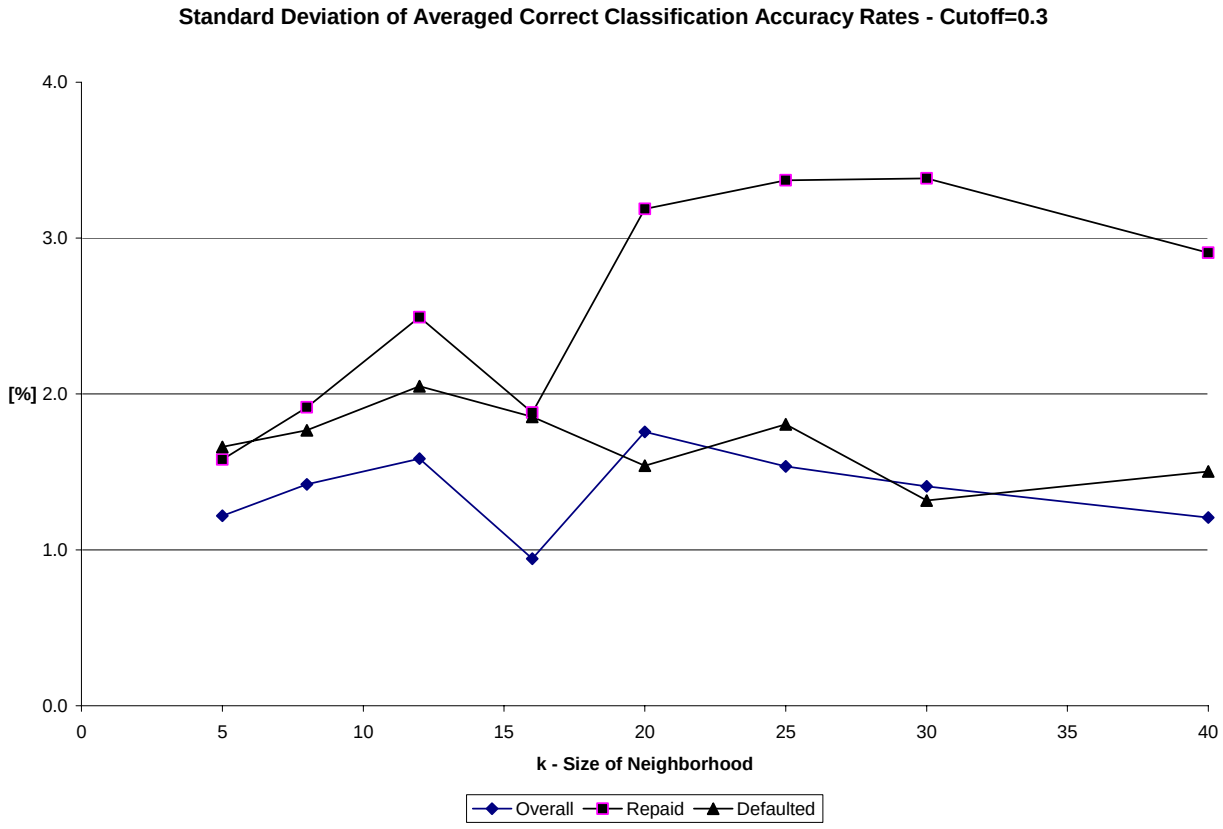

Figure 4
Standard Deviation For Averaged Overall, Repaid, And Defaulted Correct Classification Accuracy Rates [In %] For The Cutoff Probability = 0.3

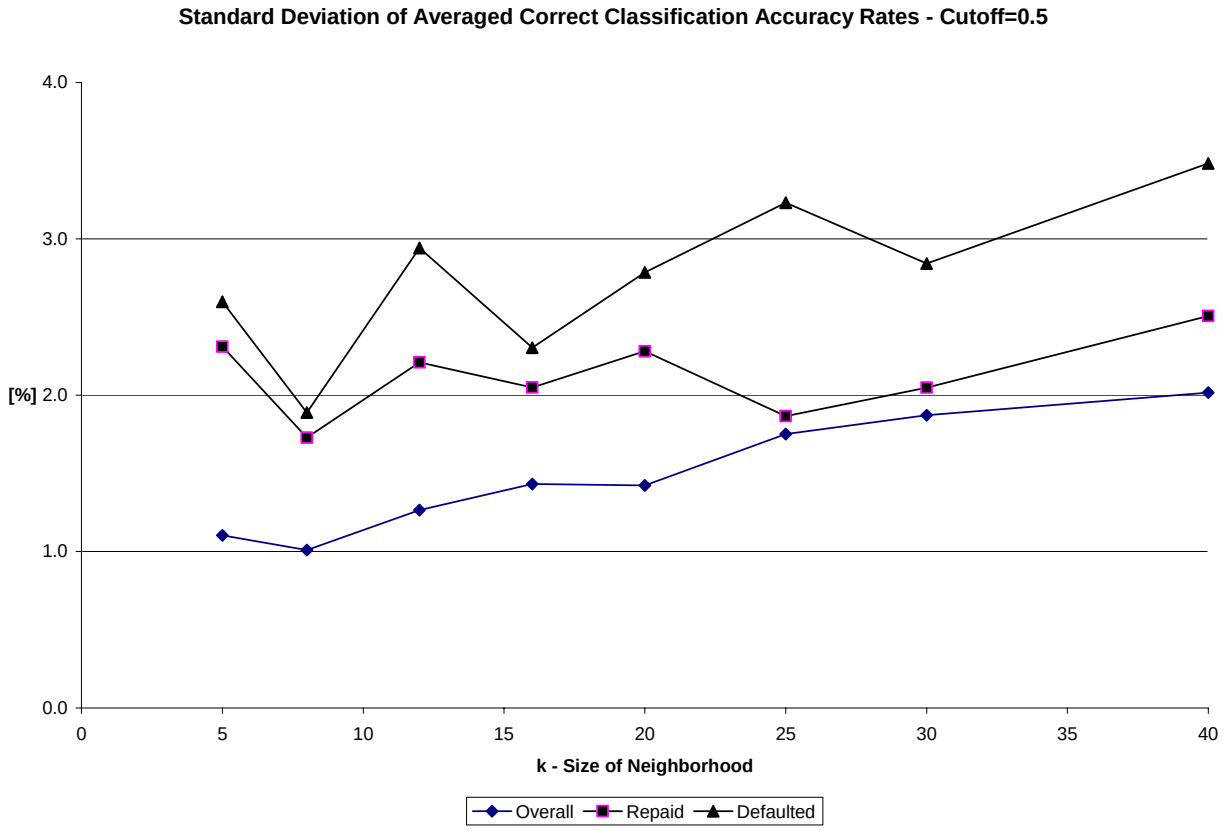

Figure 5
Standard Deviation For Averaged Overall, Repaid, And Defaulted Correct Classification Accuracy Rates [In %] For The Cutoff Probability = 0.5

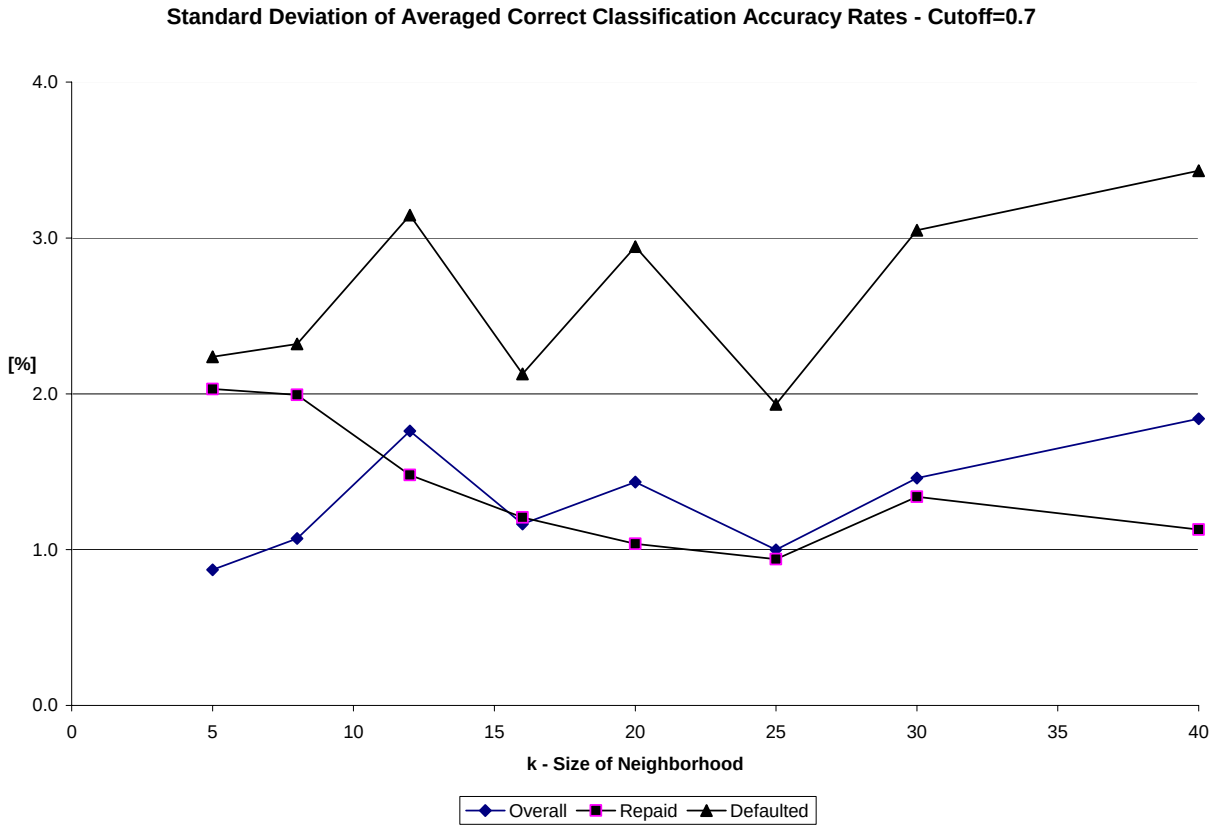

Figure 6
Standard Deviation For Averaged Overall, Repaid, And Defaulted Correct Classification Accuracy Rates [In %] For The Cutoff Probability = 0.7