Loan Application Approval using Meta Classifier

B.Rohit
Dept. of Computer Science & Engineering
JNTUH College of Engineering, Hyderabad

K.P. Supreethi
Dept. of Computer Science & Engineering
JNTUH College of Engineering, Hyderabad

ABSTRACT
A Meta Classifier in this approach is used for the approval of Loan application, as a Data mining classification tool to support Business operations in a very secure way. The goal of designing a Meta classifier system is to achieve the best possible classification performance for the task of effective decision making. This Meta classifier is the combination of Naïve Bayesian classifier, K-Nearest Neighbor and Fuzzy Set approach. This classifier focuses on combination schemes of multiple classifiers to achieve better classification performance than that obtained by individual models, which in turn is used in providing loans to the customers by verifying the various details relating to the loan such as amount of loan, lending rate, loan term, type of property, income and credit history of the customer etc. The Meta Classifier helps in analyzing the involvement of risk and behavior of the customers by distinguishing borrowers who repay loans promptly from those who don't, hence reducing the loss of revenue.

Keywords
Meta classifier, Naïve Bayesian classifier, K- Nearest Neighbor, Fuzzy set theory**.**

1. INTRODUCTION
The wide availability of huge amounts of data in banking industry and the need for transforming such data into knowledge has started realizing the need for the techniques like Data Mining which can help in analyzing huge amount of data to come up with interesting information, which can then be used throughout the organization to support the process of decision making [8]. Data mining classification techniques such as Bayesian Classification, K-Nearest neighbor classifier and Fuzzy Set approach techniques help in contributing to solve business problems by identifying patterns and trends like, how Customers will react to adjustments in interest rates, which customer is likely to accept new product offers, the risk profile of a customer segment for defaulting on loans, etc.

Classification is a form of data analysis that extracts models describing important data classes. By using classification, we can accurately predict the target class for each case in the data [1]. For example, in this implementation, a classification model is used to identify loan applicants as approved or not approved. Classification is the task that can be used to identify the class labels for instances based on a set of features or attributes. The simplest type of classification problem is binary classification. In binary classification, the target attribute has only two possible outcomes, that is, approved or not approved. Multiclass targets have more than two outcomes: for example, low, medium, high, or unknown credit rating.

The *Meta classifier* system is used to achieve the best possible classification performance for the task at hand by using the expertise of existing individual classifiers [10]. It has been observed that different classifier designs potentially offer complementary information about the pattern to be classified. The Meta Classifier in this approach is the combination of the Data Mining classification techniques, which include Naïve Bayes Classifier, K-NN Classifier and Fuzzy Set Approach. In this approach each classification technique is evaluated individually and the result of each classifier is combined with that of another in order to improve the efficiency, accuracy and credit worthiness of the result.

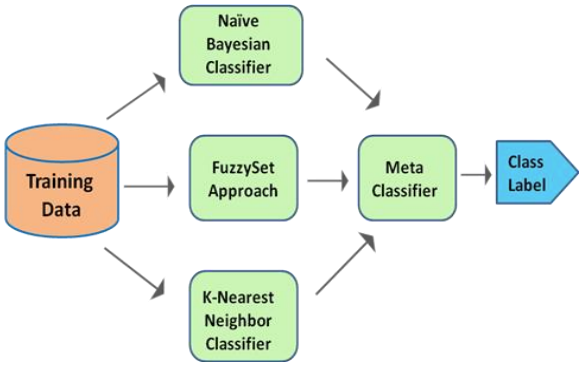


Figure 1. The Meta Classifier

2. RELATED WORK
A number of classification methods have been discussed in the literature for classification. These include naïve Bayes Classifier [5], k-Nearest Neighbor Classifier [4] and Neural Networks [2], majority voting [3], classifier combination [11], and statistical models [6], Combining Classifiers in decision trees [12].

2.1 Similarity-Based Classifier Combination for Decision Making
This study focuses on combination schemes of multiple classifiers to achieve better classification performance than that obtained by individual models, for real-world applications such as toxicity prediction of chemical compounds. The classifiers studied include voting-based k-nearest neighbors (vkNN), weighted k-nearest neighbors (wkNN), kNN model-based classifier (kNNModel) and contextual probability-based classifier (CPC). Here robust similarity-based classifier combination system is given a set of similarity-based classifiers. For this purpose, various existing classifier combination schemes and current trends in classifier combination are studied. Four previously developed classifiers are gathered to form a classifier set. The set of classifiers that were used in this study are not especially optimized for the application at hand. So different combination schemes are developed to reach reasonable classification accuracy, which is independent of the characteristic of the application [7].

2.2 Combining Classifiers with Meta Decision Trees

The paper focuses on Meta decision trees (MDTs), a novel method for combining multiple classifiers. Instead of giving a prediction, MDT leaves specify which classifier should be used to obtain a prediction. The presence of an algorithm for learning MDTs based on the C4.5 algorithm for learning ordinary decision trees (ODTs). An extensive experimental evaluation of the new algorithm is performed by combining classifiers generated by five learning algorithms: two algorithms for learning decision trees, a rule learning algorithm, a nearest neighbor algorithm and a naïve Bayes algorithm. In the experiments, performance of stacking with MDTs to the performance of stacking with ODTs is compared. The comparison between MDTs with two voting schemes and two other stacking approaches is done. Finally, the comparison among MDT's for boosting and bagging of decision trees as state of the art methods for constructing ensembles of classifiers [15].

2.3 Bayesian Classification

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. A Statistical classifier that performs probabilistic prediction, i.e., predicts class membership probabilities [1]. The Bayesian Classification is based on the following Bayes Theorem:

Bayesian Theorem: Given training data X, posteriori probability of a hypothesis H, P(H|X), follows the Bayes theorem

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})}$$

Posteriori = Likelihood X Prior/Evidence.

2.4 Naïve Bayes Classifier

A Naïve Bayes Classifier (NBC) is a simple probabilistic classifier based on Bayes rule. This is based on the simple assumption that the attribute values are conditionally independent. This assumption is called class conditional independence (i.e., no dependence relation between attributes); this greatly reduces the computation cost. It is made to simplify the computation involved and this is why it is considered "naïve" [1]. Naïve Bayes Classifier is used mainly for performing classification tasks[6].

Let D be a training set of tuples and their associated class labels, and each tuple is represented by an n-D attribute vector $\mathbf{X} = (x_1, x_2, \ldots, x_n)$. Suppose there are *m* classes $C_1, C_2, \ldots, C_m$. Classification is to derive the maximum posteriori, i.e., the maximal $P(C_i|\mathbf{X})$.

This can be derived from Bayes' theorem:

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)P(C_i)}{P(\mathbf{X})}$$

Since P(X) is constant for all classes, only $P(C_i|\mathbf{X}) = P(\mathbf{X}|C_i)P(C_i)$ needs to be maximized.

The attributes are conditionally independent of each other i.e.

$$P(\mathbf{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times \ldots \times P(x_n|C_i)$$

For each attribute, we look at whether the attribute is categorical or continuous-valued.

- If A_k is categorical, $P(x_k|C_i)$ is the no. of tuples in C_i having value x_k for A_k divided by $|C_{i,D}|$ (no. of tuples of C_i in D).
- If A_k is continous-valued, $P(x_k|C_i)$ is usually computed based on Gaussian distribution with a mean μ and standard deviation σ,

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

and $P(x_k|C_i)$ is,

$$P(\mathbf{X}|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

2.5 The k-Nearest Neighbor

K-Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it [1].The training tuples are described by n attributes and each tuple represents a point in an n-dimensional space. When given an unknown tuple, this classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "Nearest Neighbors" of the unknown tuple. Closeness" is defined in terms of a distance metric, such as Euclidean distance [4].

K-nearest-neighbor classifiersearches the pattern space for the *k* training tuples that are closest to the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, $\boldsymbol{X}_1 = (x_{11}, x_{12}, \ldots, x_{1n})$ and $\boldsymbol{X}_2 = (x_{21}, x_{22}, \ldots, x_{2n})$, is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

Min-max normalization, for example, can be used to transform a value v of a numeric attribute A to v 0 in the range [0, 1] by computing,

$$v' = \frac{v - min_A}{max_A - min_A},$$

where minA and maxA are the minimum and maximum values of attribute A.

2.6 Fuzzy Set Approach

It is also known as possibility theory, it allows us to deal with vague or inexact facts. *Fuzzy set* support a flexible sense of membership and is defined to be the pair $(x, \mu\bar{A}(x))$ where $\mu\bar{A}(x)$ could be discrete or could be described by a continuous function. The membership functions could be triangular, trapezoidal, curved or its variations. If X is a universe of discourse and x is a particular element of X, then a fuzzy set A is defined on X may be written as a collection of ordered pairs $A = \{(u, \mu_{\bar{A}}(x)), x \in X\}$[2].

3. PROPOSED METHOD

Many researchers have investigated the techniques of combining the predictions of multiple classifiers to produce a single classifier. By combining classifiers we are aiming at a

more accurate classification decision at the expense of increased complexity. The Meta Classifier is formed by the combination of Naïve Bayes Classifier, the K-Nearest Neighbor Classifier and the Fuzzy set approach. The customer data is accepted and preprocessed before applying the classification techniques. Data transformations (e.g., normalization) are applied, where data are scaled to fall within a smaller range like 0.0 to 1.0. This can improve the accuracy and efficiency of mining algorithms involving distance measurements. Then each classifier processes and evaluates the applicant data. Then the final decision for the approval of the customer loan application is done by the Meta Classifier.

3.1 Classifiers Combining Methods.

3.1.1 Voting Algorithms: Voting algorithms take the outputs of some classifiers as input and select a class which has been selected by most of the classifiers as output. We used three classifiers, Naïve Bayes Classifier, K-NN Classifier and Fuzzy Set approach. The output of each classifier is sent as input to the Meta Classifier. Then the Meta Classifier evaluates the input and processes it and generates the final result.

3.1.2 Majority Voting: If two or three classifiers agree on a class for a test document, the result of voting classifier is that class. Say, if two classifiers predict the loan to be approved for the customer based on the evaluation, then based on Majority Voting the result is considered as approved.

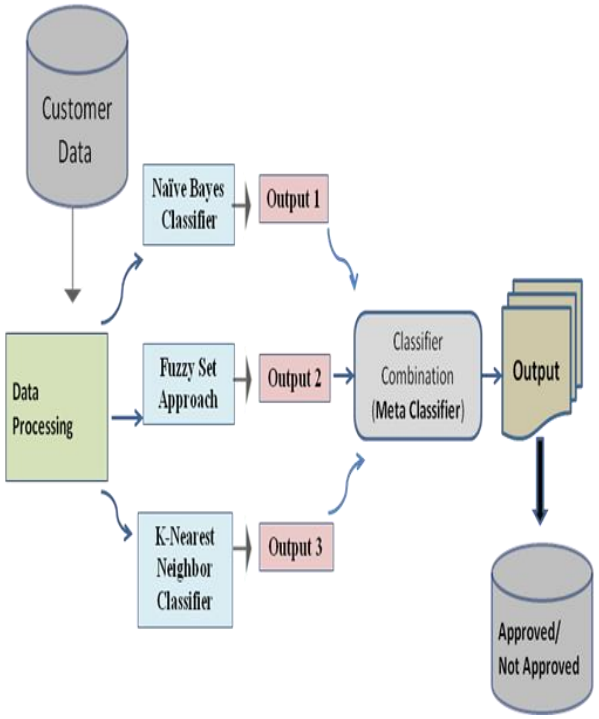


Figure 2. The Processing of the Loan Application using different Classifiers

4. IMPLEMENTATION

The Classifiers are built using Java (Business Logic) & Front End (JSP) and Oracle XE as database.

Table I
Data Set Of Loan Applicants From Lending Club.[9]

Annual Income Rs:	Is_inc_ verified	Emp length	Home Owner ship	Cibil Score
280000	Source Verified	9	rent	485
355000	Source Verified	4	rent	357
485000	Source Verified	9	mortgage	524
326000	Source Verified	7	mortgage	387
400800	Source Verified	10	own	642
505000	Not Verified	5	mortgage	741
372000	Verified	0	mortgage	354
376000	Not Verified	2	rent	457
428000	Source Verified	6	rent	651

Class:
C_1: Loan_Approved = 'yes'
C_2: Loan_Approved = 'no'
Say, the data to be classified:
X = (Income=3,20,000 Rs, Is_inc_verified= Verified, Emp_length = 5, Home_ownership= Own, Cibil_Score = 454).

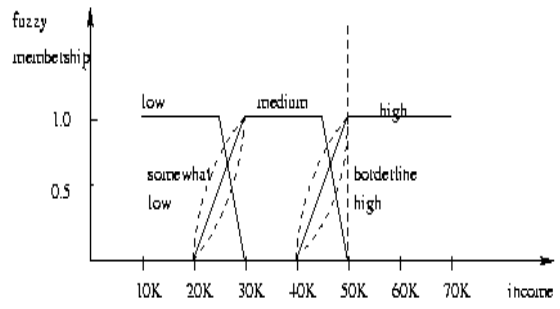
Upon Calculating the posterior probability i.e., the maximal $P(C_i|X)$ by observing the data tuple X, using *Naïve Bayes Classifier* we evaluate the data tuple and can predict that the loan applicant belongs to class "C_1" and Therefore, tuple X belongs to class ("Loan_Approved = yes").

K-nearest-neighbor classifiersearches the pattern space for thetraining tuples that are closest to the given customer data tuple, based on the search result the closeness among the tuples is calculated by using the Euclidean distance between two points or tuples.

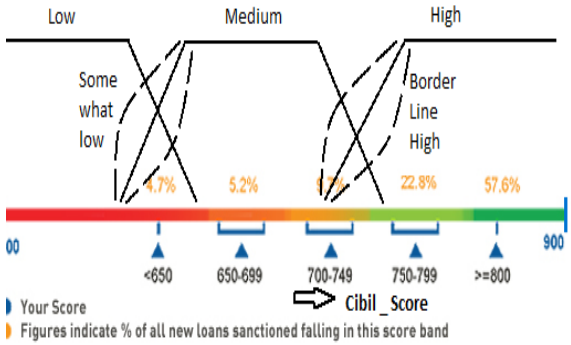
$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

Typically, we normalize the values of each attribute before, by using the above equation. This helps prevent attributes with initially large ranges (such as income) from outweighing attributes with initially smaller ranges (such as binary attributes). Min-max normalization, for example, can be used to transform a value V of a numeric attribute A to v1 in the range [0, 1]. For k-nearest-neighbor classification, the unknown tuple is assigned the most common class among its k nearest neighbors. When k = 1, the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space.

Fuzzy set logic uses truth values between 0.0 and 1.0 to represent the degree of membership (such as using fuzzy membership graph). The fuzzy membership graph includes the attribute values which are converted to fuzzy values e.g., income of the applicant is mapped into the discrete categories such as low, medium, high. The customers who are termed with high income range are likely to get loans easily than that of the customer who come in low income range.



The fuzzy membership function for Cibil Score of the applicant can be termed as either low, medium, high as follows:



The Customers with high Cibil Score (greater than 750) are likely to get the loans when compared to the customers with low Score. By evaluating the data using the fuzzy rule based system i.e., "if (Cibil_score is greater than 750) then (loan approved is equal to YES) we can generate rules.

TABLE 2.
Results of different classifiers

CLASSIFIER	ACCURACY
NAÏVE BAYES	86%
K-NN	86.5%
FUZZY SET	83.25%

The three different classifiers produce the results based on the result generated after evaluating the required data, each classifier predicts whether the loan application of the applicant can be approved or not. Voting algorithm is then applied to the results generated by the different classifiers; where in the results of the classifiers are taken as input to the Meta Classifier. The Meta Classifier then makes use of the Majority voting where the results of the classifiers are compared and evaluated. The evaluation is done based on the majority of the result generated i.e., an application must be approved by at least two classifiers to approve the loan. If any two of the classifiers predict the application as 'approved' then the Meta Classifier generates the result as 'approved' else it will generate the result as 'not approved'.

5. CONCLUSION

In this paper, we have proposed an approach for classification, which is used in providing loans to the customers by verifying the various details relating to the loan such as amount of loan, lending rate, loan term, type of property, income and credit history of the customer. Our approach is based on building a Meta Classifier by combining different classifiers. This Meta Classifier is a combination of Naïve Bayes Classifier, k-Nearest Neighbor and Fuzzy Set Approach. This Classifier makes use of voting algorithm in order to achieve a better classification rate and to improve the accuracy of the Meta Classifier while taking the decision of Loan approval process.

6. REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining concepts and techniques* 3rd ed., ISBN 978-0-12-381479-1, 2012.

[2] S. Rajasekaran, G.A. Vijayalakshmi Pai, *Neural Networks, Fuzzy Logic, and Genetic Algorithms*. ISBN-978-81-203-2186-1.

[3] M. Srinivas, K.P. Supreethi, E.V. Prasad, *Efficient text classification using best feature selection and combination of methods*. Springer BerlinHeidelberg 2009.

[4] P´adraig Cunningham and Sarah Jane Delany "*k-Nearest Neighbor Classifiers*". UCD-CSI-2007-4.

[5] Ricardas Mileris, *Estimation of loan applicants default probability applying discriminant analysis and simple bayesian classifier*, ECONOMICS AND MANAGEMENT: 2010. 15, ISSN 1822-6515.

[6] Isık Biçer1, Deniz Sevis2, Taner Bilgiç1, *Bayesian Credit Scoring Model with Integration of Expert Knowledge and customer data*. Izmir University of Economics, Turkey, 2010. ISBN 978-9955-28-598-4.

[7] Gongde Guo and Daniel Neagu "*Similarity-based Classifier Combination for Decision Making*" Member, IEEE. SMCC-05-06-0125.

[8] Rujuta Shinde, Priya Vaghurdekar, Prof. Santaji Shinde. *Delibration of Data Mining in Banking*, IJERT Vol-I Issue 8, October-2012.

[9] Loan Data Sets by "Lending Club". https://www.lendingclub.com/info/download-data.action.

[10] T.K.Ho, J.J. Hull and S. N. Srihari. "*Decision Combination in Multiple Classifier Systems*". IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 16, Issue 1, pp. 66 – 75, 1994.

[11] L. Xu, A.Krzyzak, and C. Suen. "*Methods of Combination Multiple Classifiers and Their Applications to Handwritten Recognition*". IEEE Transactions on Systems, Man and Cybernetics, SMC-22(3):418-435, May/June 1992.

[12] LJUPCˇO TODOROVSKI, SAˇSO DˇZEROSKI "*Combining Classifiers with Meta Decision Trees*" Machine Learning, 50, 223–249, 2003