COMPARATIVE ANALYSIS OF BAYES AND LAZY CLASSIFICATION ALGORITHMS

Dhyan Chandra Yadav

E-mail: dc9532105114@gmail.com
S.N.P.G. College , Narahi, Balia (U.P.)
Department of Computer Application

ABSTRACT:

Data mining applications are used in various areas such as sales, marketing, banking, finance, health care, insurance and medicine. Data Mining provide a specific environment to check two different type classifiers . The primary is Bayes Classifier and secondary is Lazy classifier , both classifiers create separate environment for consumer claim data set. In this paper we analyze the ROC performance of Bayesian and Lazy classifiers for large consumer claim data set. There are two algorithms in Bayesian classifier namely BayesNet, and Naïve Bayes. In lazy classifier has three algorithms namely IBL, IBK and Kstar.

Keywords: Bayes Classifiers: Bayes Net; Lazy Classifiers: IBK, Naïve Bayes, IBL, Kstar; Weka.

[1] INTRODUCTION

If you charge up too much on a credit card and discover that you can no longer make the payments the worst thing you can do is to stop making them. If you do this, your account could be sent to a collection agency. What’s better is to first try to trim down your expenses so that you’ll have more room to make your credit card payments. This could mean sacrificing some luxuries for a few months until you get your balance reduced. You could also contact the credit card company and ask it to lower your interest rate or minimum payment. If none of this works you may want to get help from a credit- counseling agency. When you do this, you’ll be
assigned a debt counselor who will review your debts and your earnings and help you create a budget that could make it possible for you to make those payments [1].

**BayesNet:**

K. H. Raviya and B. Gajjar proposed that Bayes Net learns Bayesian networks made in nominal attributes (numeric ones are prediscretized) and no missing values (any such values are replaced globally). Bayes Nets or Bayesian networks are graphical representation for probabilistic relationships among a set of random variables. Given a finite set \( X=\{X_1\ldots X_n\} \) of discrete random variables where each variable \( X_i \) may take values from a finite set represented by \( \text{Val}(X_i) \) [2].

![Fig.1. Tree Representation of Consumer Claim attributes by Bayes Net Algorithm](image1)

**Naïve Bayes:**

H Ian and E Frank proposed that data mining implements the probabilistic Naïve Bayes classifier. Naïve Bayes Simple uses the normal distribution to model numeric attributes. Naïve Bayes can use kernel density estimators, which develop performance if the normality assumption if grossly correct; it can also handle numeric attributes using supervised discretization. Naïve Bayes Updateable is an incremental version that processes one request at a time. It can use a kernel estimator but not discretization [3].

![Fig.2. Visualization of Instances by Naïve Bayes Algorithm](image2)
M F B Othman and T M S Yau, proposed that IBL is a basic instance-based learner which finds the training instance closest in Euclidean distance to the given test instance and predicts the same class as this training distance. If several instances qualify as the closest, the first one found is used. IBL algorithms do not construct extensional concept descriptions. Alternatively, concept descriptions are determined by how the IBL algorithm’s selected similarity and classification function use the current set of saved distances [4].

![Fig.3. Visualization of Instances by IBL Algorithm](image)

**IBK (K - Nearest Neighbour):**

B upadhyay proposed that IBK is a k-nearest-neighbour classifier that uses the same distance metric. The number of nearest neighbours can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation focus to an upper limit given by the specified value. IBK is a k-nearest-neighbour classifier. A kind of different search algorithms can be used to speed up the task of finding the nearest neighbours. A linear search is the default but further options include KD-trees, ball trees, and so-called “cover trees”. The distance function used is a parameter of the search method. The remaining thing is the same as for IBL—that is, the Euclidean distance; other options include Chebyshev, Manhattan, and Minkowski distances [5].

![Fig.4. Visualization of Instances by IBK Algorithm](image)
K -Star:

T C Sharma and M Jain proposed about K star. The K* algorithm can be defined as a method of cluster analysis which mainly aims at the partition of „n“ observation into „k“ clusters in which each observation belongs to the cluster with the nearest mean. We can describe K* algorithm as an instance based learner which uses entropy as a distance measure. The benefits are that it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing values [6].

![Fig.5. Visualization of Instances by K-STAR Algorithm](image)

[2] RELATED WORKS:

H. K analyzed that Credit cards fraudulence arises at very high level scale so we cannot easily detect and predict the related attributes but by the help of data mining classifier tool to prevent the activity of fraudsters in the misuse of credit cards uses the algorithms of neural networks. This system predicts the probability of fraud on an account by comparing the current transactions and the previous activities of each holder [7].

D C Yadav and S Pal discussed that classifier algorithms provide very accurate result in software error detection by J48, ID3 and Naïve Bayes data mining algorithms correctly classified instances will be partition in to numeric and percentage value, kappa statics, mean absolute error and root mean square error will be at numeric value only ID3 and J48 time taken to build model: 0.2 seconds and test mode :10 fold cross validation. Here Weka compare all required parameters on given instances with the classifiers respective accuracy and prediction rate based on highest accuracy of J48 is 100% without error also Naïve Bayes 100% correctly classified but with some error and ID3 95% correctly classified, so it is clear that J48 is the best in three respective algorithms so it is more accurate [8].

D C Yadav and R Kumar discussed that association algorithms provide very accurate result in the frequent and relationship between data object and find the percentage of confidence, support, of data object by the help of apriori, predictive apriori and filtered associate algorithms. Therefore these algorithms can be used in other domains to bring out interestingness among data present in the origin [9].
D C Yadav and R Kumar discussed that three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compare the performance of these three major clustering algorithms. Author compared using a clustering tool and find result: K-Means algorithm is better than Hierarchical Clustering and Make density based algorithm because all the algorithms have some ambiguity in some (noisy) data when clustered [10].

R Sukanya and K Prabha discussed that back propagation Neural Network, Support Vector Machine is used for rainfall prediction. ANN improves the efficiency of Rainfall prediction by analyzing the historical and current facts to make accurate predictions about future [11].

R S, S M, N E, S P and V Kirand discussed that the huge volume of warranty data for segregating the fraudulent warranty claims using pattern recognition and clustering. Survey of automotive industry shows up to 10% of warranty costs are related to warranty claims fraud, costing manufacturers several billions of dollars. The existing methods to detect warranty fraud are very complex and expensive as they are dealing with inaccurate and vague data, causing manufacturers to bear the excessive costs [12].

D C Yadav analyzed that in statistical analysis of binary classification, the F1 score is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. In this analysis author computed the best score for F1 by the help of data mining classifier algorithms and choose the ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has highest F1 score and take less time to build a mode [13].

D C Yadav analyzed that the Matthews correlation coefficient is used in machine learning as a measure of the quality of binary (two-class) classifiers. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. Author computed the MCC is in essence a correlation coefficient between the observed and predicted binary classifications by the help of data mining classifier algorithms and ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has highest MCC value and minimum number of time in second 0.00 to build a model [14].

D C Yadav analyzed that the informedness of a prediction method as captured by a contingency matrix is defined as the probability that the prediction method will make a correct decision as opposed to guessing and is calculated using the bookmaker algorithm. Their correlation is the generated by LAD Tree, ID3, and J48 data mining algorithms and find ID3 is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has minimum time to build a model [15].

D C Yadav analyzed that the FDR-controlling procedures provide less stringent control of Type I errors compared to class wise errors. In this analysis we choose the ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has minimum time to build a model [16].

D C Yadav analyzed that all analysis on the basis of dependable variables for overall performance and Predicts categorical class level classifiers based on training set and the values in the class level attribute use the model in classifying new data. Author analyzed between AD Tree, LAD Tree, J48 and Naïve Bayes for correctly classify and incorrectly classify with kappa static model and choose the LAD Tree is the best data mining classifier
algorithms to be applied over selected datasets. Because LAD Tree has highest correctly value 83.333% and minimum number of unclassified instances is 0.00. Also Lad tree have highest value 0.3576 of metric for accuracy [17].

T, R and Liu discussed that a framework was presented on the base of security systems and Case based reasoning for fraud detection. First, a set of normal and fraud cases are made from labeled data. Then, the primary detectors are made with random or genetic algorithms. Then, negative selection and clonal selection operations are applied on primary detectors in order to obtain a set of detectors with different algorithms that can detect a variety of frauds [18].

M, S, B and Saira discussed that many fraud detection systems that have been presented so far, have used data mining and neural network approaches. While no fraud detection system with the combination of anomaly detection, misuse detection and decision making system have been used so far for fraud detection in credit cards. Then, a system was proposed that used Hidden Markov Model to detect the fraudulent transactions [19].

A John analyzed that hybrid feature selection and anomaly detection algorithm in order to detect fraud in credit cards. The authors have noted that fraud detection on the internet must be done online and immediately. Since the use of credit card by card holders follows a fixed pattern, this fixed pattern can be extracted from a usual legal activity of card holders in 1 or 2 years. Thus, this pattern is compared to the use of process of card holder and in case of non-similarity in the pattern, the activity is considered illegal. It should be noted that the neural networks were used to teach the patterns detection in the model in this study [20].

A P, M K and AN discuss that data mining as one of the most efficient tools of data analysis has attracted the attention of many people. The use of different techniques and algorithms of this tool in various fields like customer relationship management, fraud management and detection, medical sciences, sport and etc. Due to the large number of data in banks, data mining has had lots of functions in financial and monetary affairs so far. Credit risk management, fraud detection, money laundering, customer relationship management and banking services quality management are some examples of data mining function in banks[21].

In this research work, we have analyzed the performance of Bayesian and Lazy classifiers for classifying the dataset. There are two algorithms in Bayesian classifier namely BayesNet, and Naïve Bayes. In lazy classifier has three algorithms namely IBL, IBK and Kstar. The performances of Bayesian and lazy classifiers are analyzed by applying various performance factors. From the experimental results, it is observed that the lazy classifier is more efficient than Bayesian classifier.

[3] METHODOLOGY:

Our research approach is to use Bayes Net, Naïve Bayes ,IBK,IBL and K-Star on consumer claim data set. The research methodology is divided into 5 steps to achieve the desired results:

Step 1: In this step, prepare the data and specify the source of data.
Step 2: In this step select the specific data and transform it into different format by weka.
Step 3: In this step, implement data mining algorithms and checking of all the relevant dispute is perform.

Step 4: The decision is taken on the presence of dispute in source code. If dispute is present then proceed further, otherwise it will stop. We classify the relevant dispute using Bayes Net, Naïve Bayes, IBK, IBL and K-Star.

Step 5: At the end, the results are display and evaluated.

1. Data Preparation:

If you charge up too much on a credit card and discover that you can no longer make the payments the worst thing you can do is to stop making them. If you do this, your account could be sent to a collection agency. What’s better is to first try to trim down your expenses so that you’ll have more room to make your credit card payments. This could mean sacrificing some luxuries for a few months until you get your balance reduced. You could also contact the credit card company and ask it to lower your interest rate or minimum payment. If none of this works you may want to get help from a credit counseling agency. When you do this, you’ll be assigned a debt counselor who will review your debts and your earnings and help you create a budget that could make it possible for you to make those payments. The Consumer Complaint Database is a collection of complaints on a range of consumer financial products and services, sent to companies for response. We don’t verify all the facts alleged in these complaints, but we take steps to confirm a commercial relationship between the consumer and the company.

Table 1. Representation of Computational Variables of Consumer Claim

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complaint Type</td>
<td>Consumer Complaint Database Attributes for Financial Problem (Bank, Lender &amp; Company etc.)</td>
</tr>
<tr>
<td>Sample Size</td>
<td>1000 Total: 197 Consumer dispute and 803 non dispute</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Field name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date received</td>
<td>The date the CFPB received the complaint.</td>
</tr>
<tr>
<td>Tags</td>
<td>Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers.</td>
</tr>
<tr>
<td>Date sent to company</td>
<td>The date the CFPB sent the complaint to the company.</td>
</tr>
<tr>
<td>Company response to consumer</td>
<td>This is how the company responded. For example, “Closed with explanation.”</td>
</tr>
<tr>
<td>Timely response?</td>
<td>Whether the company gave a timely response. For example, “Yes” or “No.”</td>
</tr>
</tbody>
</table>
2. Data Selection and transform:

The database generally updates daily, and contains certain information for each complaint, including the source of the complaint, the date of submission, and the company the complaint was sent to for response. The database also includes information about the actions taken by the company in response to the complaint, such as, whether the company’s response was timely and how the company responded. If the consumer opts to share it and after we take steps to remove personal information, we publish the consumer’s description of what happened. Companies also have the option to select a public response. Company level information should be considered in context of company size. If you identify an error on your credit report, you should start by disputing that information with the credit reporting company. You should explain in writing what you think is wrong, why, and include copies of documents that support your dispute.

After data selection we transform data by data mining weka tool. These tree algorithms have his specific details as accuracy by class ,Time to build a model and stratified cross validation with summary. Confusion matrix describe correctly classified and incorrectly classified instances with respect to class

3. Implementation:

Classification is an important data mining technique with broad applications. It is used to classify each item in a set of data into one of predefined set of classes or groups. Classification algorithm plays an important role in document classification. In this research, we have analysed two classifiers namely Bayesian and lazy. In Bayesian classifier, we have analysed two classification algorithms namely BayesNet and naïve bayes, in lazy classifier we have analysed three classification algorithms such as IBL, IBK and Kstar. Lazy learners store the training instances and do no real work until classification time. Lazy learning is a learning method in which generalization beyond the training data is delayed until a query is made to the system where the system tries to generalize the training data before receiving queries.

[4] RESULT AND DISCUSSION:

Table.2. Representation of Computation of Instances Part-I

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy</th>
<th>Kappa</th>
<th>RMSE</th>
<th>RAE</th>
<th>RRSE</th>
<th>TIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Net</td>
<td>66.6</td>
<td>0.14</td>
<td>0.38</td>
<td>119.50</td>
<td>115.71</td>
<td>0.02</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>73</td>
<td>0.10</td>
<td>0.35</td>
<td>110.23</td>
<td>106.48</td>
<td>0.00</td>
</tr>
<tr>
<td>IBK</td>
<td>78.2</td>
<td>0.06</td>
<td>0.35</td>
<td>89.47</td>
<td>106.81</td>
<td>0.00</td>
</tr>
<tr>
<td>IBL</td>
<td>76.7</td>
<td>0.05</td>
<td>0.34</td>
<td>106.33</td>
<td>103.39</td>
<td>0.05</td>
</tr>
<tr>
<td>K Star</td>
<td>78.3</td>
<td>0.08</td>
<td>0.34</td>
<td>91.80</td>
<td>103.36</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table.3. Representation of Computation of Instances Part-II
The following tables show the accuracy measure of classification techniques. They are the True Positive rate, F Measure, Receiver Operating Characteristics (ROC) Area and Kappa Statistics. The TP Rate is the ratio of play cases predicted correctly cases to the total of positive cases. It is a probability corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement.

From the all observation of Table.2. and Table.3. we find that:

- Classification accuracy value 78.3 of K-Star is greater than other algorithms.
- RMSE error value 0.34 of K-Star is not bigger than other algorithms.
- RAE error value 91.80 of K-Star is not bigger than other algorithms.
- RRSE error value 103.39 of K-Star is not bigger than other algorithms.
- Time build model 0.00 Sec. of K-Star is not bigger than other algorithms.
- ROC value 0.60 of K-Star is not less compare to other algorithms.

[5] CONCLUSION

Data mining can be defined as the extraction of useful knowledge from large data repositories. In this paper, the classification algorithms namely Bayesian and Lazy classifier are used for classifying consumer claim data set which are stored as csv file. The Bayesian Algorithm includes two techniques namely Bayes Net, Naïve Bayes and the Lazy algorithms includes IBl (Instance Based Learning), IBK (K-Nearest Neighbour) and KStar techniques. By analyzing the experimental results it is observed that the lazy classifier of K-Star classification technique has a better result compared to other techniques.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>T P Rate</th>
<th>F P Rate</th>
<th>Precision</th>
<th>F-Measure</th>
<th>ROC</th>
<th>CLASS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayes Net</td>
<td>0.41</td>
<td>0.26</td>
<td>0.26</td>
<td>0.32</td>
<td>0.59</td>
<td>YES</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.28</td>
<td>0.13</td>
<td>0.25</td>
<td>0.22</td>
<td>0.60</td>
<td>YES</td>
</tr>
<tr>
<td>IBK</td>
<td>0.04</td>
<td>0.04</td>
<td>0.21</td>
<td>0.07</td>
<td>0.61</td>
<td>YES</td>
</tr>
<tr>
<td>IBL</td>
<td>0.06</td>
<td>0.05</td>
<td>0.22</td>
<td>0.10</td>
<td>0.59</td>
<td>YES</td>
</tr>
<tr>
<td>K Star</td>
<td>0.05</td>
<td>0.04</td>
<td>0.22</td>
<td>0.09</td>
<td>0.60</td>
<td>YES</td>
</tr>
</tbody>
</table>
REFERENCES


