MULTIFACTOR NAÏVE BAYES CLASSIFICATION FOR THE SLOW LEARNER PREDICTION OVER MULTICLASS STUDENT DATASET

Swati¹, Rajinder Kaur²

¹M.E student of Chandigarh University, Gharuan
²Assist. Professor of Chandigarh University, Gharuan

ABSTRACT

The high school students must be observed for their slow learning or quick learning abilities to provide them with the best education practices. Such analysis can be perfectly performed over the student performance data. The high school student data has been obtained from the schools from the various regions in Punjab, a pivotal state of India. The complete student data and the selective data of almost 1300 students obtained from one school in the regions has been undergone the test using the proposed model in this paper. The proposed model is based upon the naïve bayes classification model for the data classification using the multi-factor features obtained from the input dataset. The subject groups have been divided into the two primary groups: difficult and normal. The classification algorithm has been applied individually over data grouped in the various subject groups. Both of the early stage classification events have produced the almost similar results, whereas the results obtained from the classification events over the averaging factors and the floating factors told the different story than the early stage classification. The proposed model results have shown that the deep analysis of the data tells the in-depth facts from the input data. The proposed model can be considered as the effective classification model when evaluated from the results described in the earlier sections.

KEYWORDS

Slow learner prediction, data classification, averaging factor classification, floating factor classification.

1. INTRODUCTION

Self-organizing Sensor networks dynamically changes the network topology and distributed either randomly or uniformly. A huge amount of tiny sensor nodes (SNs) monitor temperature, humidity, motions and sound. In multi-hop transmission of WSN each sensor nodes play dual characteristic of perceiving the environment and forwards the collected data to the base station (BS) via integrated radio transmitters. The key challenge is to prolong the lifetime of WSN since it is not possible to recharge the batteries of SNs in unattended environment. Considering every node in the network for a time periodic data collection generates more traffic. So the period for data collection is to be enough for collecting data from nodes. To avoid traffic congestion and packet drop over transmission only random nodes to be selected for data collection in every miniature period. Therefore, energy efficient mechanisms are required for computation operations like data storage, path construction and decision making of source nodes and to secure the communication from sources to sink.

In the Internet, web is a vast, dynamic, diverse and amorphous data repository that stores information/data in incredible amount and also enhances the complexity to deal with the information from different opinion of users, view, and business analyst and web service providers. The Internet service providers desire to search the technique to guess the user’s behaviors and customize information to shrink the traffic load and create the Web site suited for...
different set of users. The business analysts need to have tools to learn consumer's needs. All of
them are expecting equipment or techniques to help them satisfy their requirements and solve the
problems encountered on Web. Thus, Web mining has become trendy active area and is taken as
the research topic for this analysis.

In the prediction model, we find out what users are looking for on internet or Few user so be
might be survey at only documented data. It is the submission of form and facts of mining
techniques to find out interesting usage pattern from World Wide Web facts and figures in the
alignment to realize and better serve the desires of Web based applications. Usage figures and
facts hold the personal or source of World Wide Web users along with browsing at World Wide
Web site. Web usage excavation itself could be categorized farther counting on the kind of usage
facts and figures considered:

- Web Server Data: The client logs are anthologized by the Web server.

1.1 Application Server Data

Financial submission servers have significant characteristics to endue ecommerce
submissions to be built on peak of them with tiny effort. A key feature is the dexterity to
pathway diverse kind of enterprise events and logs them in application server logs.

1.2 Application Level Data

New type of events can be characterized in an application, and logging can be two times on
for them, therefore generating histories of these particularly characterized events. It should be
noted however, that numerous end submissions need a combination of one or more then one
of the methods directed in the classes above.

In the current scenario of the World Wide Web, the popularity is increasing day by day, so as the
web mining. In the web, increasing the number of websites and web users, the data of web usage
is stored on web servers. By analysis of web server data, we have several information such as
user surfing behavior that is most crucial aspect of business marketing which helps to user profile,
web site designs meliorate and make better marketing and business decision making website user
friendly and popular. With looking at minimum data we cannot identify patterns, for purpose of
analyst need significantly huge amount of data. All the data collected by the service providers is
stored in the high capacity servers. As the user becomes high in the numbers this data also grows
and the data logs are not easy to maintain. To analyze such kind of large data we need to extract
the useful data and this data is then mined to get the patterns of the user behavior. For this
purpose an efficient algorithm is needed, which can do the purpose and help extracting the
information. There are so many algorithms which are liable to resolve the purpose but they all
take the time in scanning and pattern matching. In this synopsis, an algorithm is designed which
employs the website architecture and gives the information about the users’ usage behavior. The
users access a website by going from one page to another page, by the hyperlinks provided in the
web pages. Mining the information identified by the analysis, will not only help making the user
interface better but also in various business decision making.

The user traverses web-site in different-different ways. The variations between traversal patterns
increase the complexity of obtained information from path traversed. There are several available
algorithms for citing the user traversal patterns. In this paper a new approach has been proposed
for mining the large reference patterns. The traversal patterns have achieved by first mining the
maximal forward references take away the web server log and after this maximal forward
reference are obtained and large references can be calculated that are most frequently used paths
followed by user for website.
2. **Literature Review**

Eltahir, Mirghani, and Anour FA Dafa-Alla [2] have proposed to extract the information from web server logs using prediction model. The major problem which faces any website admin or web application system is data increase per-second that is stored in different formats and types in server log files about users, future needs and maintains structure and apprised of website or web services according to their preceding data. Prediction model aims at discovering useful information or learning from usage data registered in log files, based on primary kinds of data used in the mining process. By using one of the web mining techniques, this paper cause a prediction model techniques to procure knowledge from web server log files where all user piloting history is registered. Gupta, Ashika, Rakhi Arora, Ranjana Sikarwar, and Neha Saxena [3] have projected a technique for prediction model using improved Frequent Pattern Tree algorithms. Prediction model itself can be categorized further depends upon usage data considered are application server, web server and application level data. This Research work target on web use mining and especially keeps tabs on running crosswise the web utilization examples of sites from the server log records. The binding of memory and time usage is compared by means of Apriori algorithm and refined Frequent Pattern Tree algorithm. Sharma, Murli Manohar, and Anju Bala [4] discussed an algorithm for frequent access pattern identification in prediction model. In web mining the analysis of web logs is done to identify the user search patterns. In general approaches of find the patterns, pattern tree is created and the analysis is done, but in proposed algorithm there is no need of tree creation and the analysis is done based on the website architecture, which will increase the ability of the other pattern matching algorithms and needs only one database scan. Bhargav, Anshul, and MunishBhargav [5] have worked on pattern discovery and users classification through prediction model. The proposed structure is based on three steps. In the first step, preprocessing is done to remove useless data from web log file so as to depreciate its size. In the second step, this clean up the log file is used for discovering usage patterns. For ever, the discovered patterns conduct to the classification of users: on the basis of countries; on the basis of direct portal to the site or associated by the new site; on the basis of time of connection , i.e., either different seasons or different months or peculiar days. This information can then be used by the website administrators for efficient legislation and personal of their websites and thus the specific needs of express communities of users can be fulfilled and so the profit can be increased.

3. **Simulation Model**

Prediction is making claims about the something that will happen, often based on information from past and from current state. Everyone has solved their issue of prediction every day with several degrees of success. For example harvest, weather, energy consumption, movements of fore x currency pairs or of shares of stocks, earthquakes, and lot of stuff needs to be predicted. In technical domain predictable of system can be often expressed and evaluated using equations - prediction is simply evaluation or solution of equations. However, practically face the problems where such a description would be too complicated or not possible fully. In addition, the solution by method could be complicated computationally, and sometimes get the solution after event to be predicted happened. There is possibility to use several approximations, for example regression of dependency of predicted variable on events which is extrapolated to future. Find such approximation can be also difficult. This approach is generally means create the model of predicted event. Neural networks can be used for prediction with several levels of success. The advantage of automatic learning of dependencies only from measured the data without any need to add further information. The neural network is trained from the historical data with the hope that it will be discover hidden dependencies and that it will be able to use them for predicting into future. In other words, neural network is not represent by an explicitly given model. It is more a black box that is able to learn something. It is possible to be predict various types of data, however in the rest of this text we will focus on predicting of time series. Time series show the
development of a value in time. Of course, the value can be impact by also other factors than just for time. Time series are be represents discrete history of a value and from a continuous function it can be obtained using sampling.

The Bayesian Classification is represented supervised learning method too a statistical method for classification. Assume an underlying probabilistic model which allowscapturing uncertainty about the model in a principle way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

![Bayesian Classification Diagram](image)

In this Classification is named after Thomas Bayes (1702-1761), who projected Bayes Theorem. Bayesian classification offers prior knowledge and practical learning algorithms and observed data can be combined. explicit probabilities to hypothesis and it is robust to noise in input data.

The slow-learner prediction model is used for purpose of slow learner prediction model using adaptive classification model. Depends upon the precise nature of probability model, naïve Bayes classifiers can be trained efficiently in the supervised learning setting. Naive Bayes classifiers work much better in several complex situations than one might expect. Here independent or dependent variables are considered for the purpose of the prediction or existence of the crisis. In spite of their naïve design oversimplified assumptions, naïve Bayes classifiers often work much better in more complex real world situations and it is also solve the floating point values problems. Recently, careful analysis of the Bayesian classification problem has shown the some theoretical reasons of the apparently unreasonable efficacy of naïve Bayes classifiers. An advantage of naïve Bayes classifier is that it required a small amount of training data to estimate the parameters that are necessary for classification.
Navie Bayes Classifier Algorithm is generic is described as following:

1. Each data sample is represented by an n dimensional element vector, $X = (x_1, x_2, \ldots, x_n)$, depicting n measurements made on the sample from n attributes, respectively $A_1, A_2, A_n$.
2. Suppose that there is m classes, $C_1, C_2, \ldots, C_m$. Given an unknown data sample, $X$ (i.e., having no class label), the classifier desire predict that $X$ belongs to the class having the highest posterior probability, conditioned on $X$. That is, the naive probability appoint an unknown sample $X$ to the class $C_i$.
3. if and only if:
   $$P(C_i/X) > P(C_j/X) \text{ for all } 1 \leq j \leq m \text{ and } j \neq i$$
   So we maximize $P(C_iX)$. The class $C_i$ for which $P(C_iX)$ is maximized is called the maximum posteriori hypothesis. Beyond Bayes theorem,
   $$P(C_i/X) = \frac{P(X/C_i)P(C_i)}{P(X)}$$
   As $P(X)$ is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class anterior probabilities are not known, then it is commonly assumed that the classes are same likely, i.e.,
   $$P(C_1) = P(C_2) = \ldots = P(C_m),$$
   and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$.

Algorithm 1: Naive Bayes Classifier for Slow Learner Prediction

- Computation diagnosis="yes", diagnosis="no" probabilities $P_{yes}$, $P_{no}$ from training input.
- For Each Test Input Record
- For Each Attribute
- Count Category of Attribute Based On Categorical Division
- Calculate Probabilities Of Diagnosis=“Yes”, Diagnosis=“No” Corresponds To This Category $P(Attr,Yes)$, $P(Attr,No)$ From Training Input.
- For Each Attribute
- CountResultyes= Resultyes* $P(Attr,Yes)$, Resultno= Resultno* $P(Attr,No)$;
- Calculate Resultyes= Resultyes * $P_{yes}$
- Resultno= Resultno* $P_{no}$;
- If(Resultyes>Resultno) So Diagnosis="Yes";
- Else Then Diagnosis="No";

The Formulae used under the Naïve Bayes classifier algorithm:

- $P_{yes}$=total number of yes/total no. of records.
- $P_{no}$=total number of no/total no. of records.
- $P(\text{attr,} \text{yes})$=total number of yes in corresponding category/entire number of yes.
- $P(\text{attr,} \text{no})$=total number of yes in corresponding category/entire number of yes.

Algorithm 2: Customized Naïve bayes classification model

1. Load the student review dataset
2. Select the parameter set according to the input requirement
3. Select the naïve bayes for the classification model
4. Run the classification model
5. Initiate the iteration parameters
6. For each input record
   a. For each attribute
i. The entities are divided in the separate categories according the categorical data.
ii. The probability is calculated from the training input
7. For each attributes
   a. Calculate the probability and classify the data according the found irrelevance parametric setup.
   b. Return the diagnosis parameters
8. Return the sentiment classification data

4. RESULT ANALYSIS

4.1 Precision:
Precision can be defined as the ratio of related retrieved documents and the information needed by the users. High precision defines this the algorithm returns results that are relevant as compared to irrelevant results. It also defines a predictive value that is positive and this is defined in terms of the binary classification. This classification defines the documents that are retrieved. It is defined in terms of the results that the system returns at some close-off rank. Precision is also known as sensitivity.

\[
\text{Precision} = \frac{A}{A + D}
\]
Where, A depicts True Positive, B gives the True Negative, C depicts the False Negative and D is the False Positive.

4.2 Recall
Recall is the probability that a test will indicate 'test' amid those with the matching sample.

\[
\text{Recall} = \frac{A}{A + C} \times 100
\]

4.3 True Positive Rate (TPR)
True positive rate is describe as division of system whichdoes not matches patterns of input correctly to template that is non matching. It is defined percentage of inputs that are valid. True positive rate is dependent upon threshold. It is also defined as the measure that an attempt by the user that is unauthorized will be accepted by the classification calssic.

4.4 False Positive Rate (FPR)
False positive rate is describe as the probability of a system to detect the matching between the pattern that is given as input and the matching template. It is the fraction of number of false appearances to the number of attempts that are identified. It defines a measure that an attempt by the user that is unauthorized is rejected by the classification model.

Table 4.1: The classification instances based analysis

<table>
<thead>
<tr>
<th>CLASS</th>
<th>NUMBER OF INSTANCES</th>
<th>PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Instances</td>
<td>Classified 10279</td>
<td>92.98%</td>
</tr>
<tr>
<td>Incorrectly Instances</td>
<td>Classified 776</td>
<td>7.02%</td>
</tr>
</tbody>
</table>
The correctly classified and incorrectly classified instances along with the percentage has been studied in the table 4.1. The table shows the 10279 samples as the correctly classified instances, which makes the 92.98% percent of the total instance population, whereas the 776 samples have been incorrectly classified using the proposed model and it makes the 7.02% of the total instances.

![Percentage of Classification Instances](image1)

**Figure 4.1: Percentage of classification instances**

The figure 4.1 shows the graphical presentation of the table 4.1 in the form of the bar graph. The left side bar graph shows the higher level of percentage, whereas the incorrect bar in the right side is showing very low level which shows the marginal value of the incorrectly classified samples using the naïve bayes classifier over the direct score based evaluation.

![No. of Classification Instances](image2)

**Figure 4.2: Number of classification instances**

The figure 4.1 shows the graphical presentation of the table 4.1 to depict the instance classification in the form of number of instances. The left side bar graph shows the higher number of correctly classified instances, whereas the right side bar shows the incorrect number of instances using the naïve bayes classifier over the direct score based evaluation.
The table 4.2 shows the study of various performance errors, which are calculated using the statistical measures, which are studied in the form of the statistical error type 1 and type 2. The statistical error type 1 and type 2 have been measured in the performance error wheel. The highest error has been recorded for the root relative squared error, which shows the difference between the multi-method-based evaluation of the proposed model results.

### Table 4.2: The results of the statistical cross-validation tests

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kappa Statistics</td>
<td>0.86%</td>
</tr>
<tr>
<td>Mean Absolute Error</td>
<td>0.089%</td>
</tr>
<tr>
<td>Root Mean Squared Error</td>
<td>0.23%</td>
</tr>
<tr>
<td>Relative Absolute Error</td>
<td>18.12%</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>46.39%</td>
</tr>
<tr>
<td>Total Number of Instances</td>
<td>11055</td>
</tr>
</tbody>
</table>

The table 4.2 shows the study of various performance errors, which are calculated using the statistical measures, which are studied in the form of the statistical error type 1 and type 2. The statistical error type 1 and type 2 have been utilized to read the various performance errors.

![Statistical Error Analysis](image)

**Figure 4.3: Statistics Error Analysis of the classification model**

The table 4.3 shows the graphical presentation of the study of various performance errors, which are calculated using the statistical measures in the above table 4.2. The density of the errors has been measured in the performance error wheel. The highest error has been recorded for the root relative squared error, which shows the difference between the multi-method-based evaluation of the proposed model results.

### Table 4.3: The parametric results obtained from the simulation

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>CLASS</th>
<th>WEIGHTED AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>True Positive Rate</td>
<td>0.904</td>
<td>0.95</td>
</tr>
<tr>
<td>False Positive Rate</td>
<td>0.05</td>
<td>0.096</td>
</tr>
<tr>
<td>Precision</td>
<td>0.936</td>
<td>0.926</td>
</tr>
<tr>
<td>Recall</td>
<td>0.904</td>
<td>0.95</td>
</tr>
<tr>
<td>F-Measure</td>
<td>0.919</td>
<td>0.938</td>
</tr>
<tr>
<td>ROC Area</td>
<td>0.981</td>
<td>0.981</td>
</tr>
</tbody>
</table>

The table 4.3 shows the study of various performance parameters, which are calculated using the statistical measures, which are studied in the form of the statistical error type 1 and type 2. The statistical measures have been utilized to read the various performance parameters.
Table 4.4: The overall result confusion matrix

<table>
<thead>
<tr>
<th>CLASSIFIED AS</th>
<th>A = -1</th>
<th>B = 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>4427</td>
<td>305</td>
</tr>
<tr>
<td>B</td>
<td>471</td>
<td>5852</td>
</tr>
</tbody>
</table>

The table 4.4 shows the measurement of the number of instances according to the statistical type 1 and type 2 errors, which are calculated using the statistical measures, which are studied in the form of the statistical error type 1 and type 2. The statistical measures have been utilized to read the various performance parameters. The proposed model results have been obtained in the form of various performance parameters.

Figure 4.4: Naïve based classification based upon the student group 1

The results of the first method of the proposed model have been shown in the figure 4.4. The naïve bayes classification has been performed over the input data using the subject group one. The subject group one includes the subjects of mathematics and science, which are considered as the higher order classification of the slow learner students out of the given databases. The most normalized classification spread has been shown in the figure 4.4 and 4.5. The figure 4.5 shows the classification results over the subject group 2, which includes the social studies and the English language. The proposed model has performed almost similar in the case of both of the subject groups.

Table 4.5: The pre-classification categorization results

<table>
<thead>
<tr>
<th>Value</th>
<th>Count of Values</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>1155</td>
<td>88.85</td>
</tr>
<tr>
<td>Slow</td>
<td>145</td>
<td>11.15</td>
</tr>
</tbody>
</table>
The table 4.5 contains the results obtained from one entity (or one school), which contains the total records of the 1300 students from the 8th, 9th and 10th grades. The earlier stage manual classification shows the results shown in the table 4.5, where the 88.85% students are considered as the normal students and the others are estimated as the slow learners.

![Figure 4.5: Naïve based classification based upon the student group 2](image)

<table>
<thead>
<tr>
<th>Category</th>
<th>normal</th>
<th>Slow</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Stage Classification</td>
<td>88.85</td>
<td>11.15</td>
</tr>
<tr>
<td>Subject Class 1</td>
<td>40.8</td>
<td>10.54</td>
</tr>
<tr>
<td>Subject Class 2</td>
<td>47.2</td>
<td>13.14</td>
</tr>
<tr>
<td>Averaging Factors</td>
<td>44.82</td>
<td>6.3</td>
</tr>
<tr>
<td>Floating Factors</td>
<td>11.42</td>
<td>3.71</td>
</tr>
</tbody>
</table>

The 4.6 shows the results obtained from all of the classification stages for the prediction of the slow learner students out of the given data.

5. CONCLUSION

The proposed model has been programmed to return the results in the form of accuracy and class density along with the classification obtained from the input dataset. The proposed model has been designed to detect and classify the input dataset to the given classifier by using the results obtained from the online student data portal. The experimental results have been obtained from the simulation model in the form of the enlisted performance parameters. The proposed model output has been designed in the way to perform all of the operations in the sequential order as per the system design. The simulation model detects the abnormalities in the given API data. The
The proposed model has been designed to detect and track the slow learners in the given database which is observed on the basis of the input dataset. The proposed model results show the effectiveness in the automatic classification of the student data. The proposed model has performed better for the classification of the student data.

REFERENCES


AUTHOR

Swati is completed the B.Tech in 2014. She is pursuing the Master of Engineering in the Chandigarh university Gharuan, Kharar from the 2014 to till date. Her area of interest in the Education data mining is sub area of Data mining.