Performance Analysis of various Data mining Algorithms in Educational Domain Datasets

Nandini N

Abstract— Educational data mining applications are widely accepted now a day as they will help in analyzing and predicting information’s useful for enhancing educational growth. One of the major applications of this kind is the prediction of student performance in higher education. This will help the stakeholders to understand the effect of various factors in academic performance thereby enabling them to take immediate and adequate remedial actions. This research aims to understand the various attributes and their impact on the students’ academic performance. A synthetic dataset is chosen to experiment with the various data mining algorithms. Further a real time data set collected from a high school is also experimented with similar algorithms.

Index Terms— Clustering, correlation, data mining algorithms, educational data mining, KStar, PART, WEKA.

1 INTRODUCTION

In the current scenario where everyone uses various electronic devices to collect data for variety of applications including monitoring and surveillance for security and health reasons, the amount of data generated becomes quite huge. Extracting the required information from this voluminous data is quite difficult. Here comes the importance of data mining algorithms. Data mining technology finds application in all fields wherever huge data is involved. Educational Data Mining (EDM) is the utilization of Data Mining strategies on instructive information. The target of EDM is to break down such information and to determine instructive research issues. EDM manages growing new strategies to investigate the instructive information, and utilizing Data Mining techniques to more readily comprehend understudy learning condition The EDM procedure changes over crude information originating from instructive frameworks into helpful data that might greatly affect instructive research and practice. Educational data mining employs a great role in enhancing the various aspects in educational domain say from predicting the students’ performance and instructor excellence or even for administrative enhancements and proper resource utilization [1]. There are many areas where Educational Data Mining are used like Analysis and perception of information, Providing criticism for supporting educators, Recommendations for understudies, Predicting understudy execution, Student displaying, Detecting bothersome understudy practices, Grouping understudies, Social system investigation, Developing idea maps, Constructing courseware and Planning and booking[2].

2 CLASSIFICATION ALGORITHMS IN DATA MINING

Various classification algorithms are used to test the data set they are explained briefly in this section.

2.1 Multilayer Perceptron

A multilayer perceptron (MLP) is a neural system interfacing numerous layers in a coordinated diagram, which implies that the sign way through the nodes just goes one way. Every node, aside from the info nodes, has a nonlinear enactment work. A MLP utilizes back propagation as a managed learning system. Since there are numerous layers of neurons, MLP is a profound learning strategy. MLP is generally utilized for taking care of issues that require regulated learning just as examination into computational neuroscience and parallel appropriated preparing. Applications incorporate discourse acknowledgment, picture acknowledgment and machine interpretation [3].

2.2 Naïve Bayes 00

The Naive Bayesian classifier depends on Bayes’ hypothesis with the autonomy presumptions between indicators. A Naive Bayesian model is anything but difficult to work, with no confounded iterative parameter estimation which makes it especially helpful for exceptionally enormous datasets. In spite of its effortlessness, the Naive Bayesian classifier regularly does shockingly well and is broadly utilized in light of the fact that it frequently beats increasingly refined characterization techniques [4].

2.3 KStar

This is an occurrence based classifier that is the class of a test case depends on the class of those preparation occasions like it, as dictated by some similitude work. It varies from other case based students in that it utilizes an entropy-based separation work [5].

2.4 PART

It is a different and-vanquish rule student. The calculation creating sets of rules called „decision lists“ which are arranged arrangement of rules. Another information is contrasted with each standard in the rundown thusly, and the thing is allocated the class of the main coordinating guideline. PART manufactures an incomplete choice tree in every cycle and makes the "best" leaf into a standard [6].

2.5 J48

J48 is an expansion of ID3. The extra highlights of J48 are representing missing qualities, choice trees pruning, constant property estimation ranges, inference of rules, and so on. In the WEKA information mining instrument, J48 is an open source Java execution of the C4.5 calculation [7].
The rest of the paper is organized as follows. Section 2 gives the overview of the related works, Section 3 mention about the tools and data set used. Section 4 describes the experimental results obtained followed by conclusions and future scope in Section 5. References are also mentioned.

3 RELATED WORKS

Educational data mining is quite important now a day as it helps to predict the exact results will helps in planning and improvement. Many researches is happening in this area which helps pupil to improve their scores or focus more on their areas of interest or to explore more on what they are good at. Kaur et al. had experimented on a real time dataset from a high school. They used data mining prediction and classification algorithms on this real time data set and was able to analyze the student performance and predict the slow learners [1]. A similar work in Malaysian context was performed by Shahiri et al. [8]. According to the authors the attributes selection for performance analysis is very important. They had given detailed information regarding the same along with the various algorithms suitable for effective performance prediction. Educational data mining is not related to student level achievement predictions or performance analysis alone rather it can be used for instructor and administrative level performance analysis also. Such an analysis using different classifier models for evaluating the success of instructor was done in [9]. According to the authors in [10] educational data mining can do a lot for the overall improvement of the educational institutions. It can be used to assess student performance for increasing passing rates, improve the institution performance, optimizing the resource utilization and even for curriculum updates. They had given an overview of the different effective educational data mining methodologies. In [11] an organized review of the data mining algorithms in educational data mining was done. The use of clustering as a pre-processing step before the application of data mining algorithms was emphasized here. The student background and social activities also have an impact on their academic success. This was proved in [12] where the authors analyzed these attributes and the relationship with academic success. The use of co-training semi supervised learning in order to predict the undergraduate student performance was done by the authors in [13]. For this study the student characteristics and academic achievements and also the involvement in online tutorials are considered. A similar study was conducted by Polyzou and Karypis [14] wherein the focus was to more accurately predict the poor performers. They had used specific attributes which will add accuracy for the predicted result.

4 METHODOLOGY

WEKA is an assortment of AI calculations for data mining errands. The algorithms available in the WEKA can either be applied straightforwardly to a dataset or called from your own Java code. There are some tools available in WEKA for data like pre-processing, clustering, regression. Etc. [15].The data set is downloaded from report builder the link is given below [16].

https://wpreportbuilder.com/examples/students-exam-marks-list-generate-excel-xlsx/ . The student report cards were collected from a high school in a rural area of Tamilnadu. The various characteristics influencing the students' academic performance were taken as the input. It is not a single factor that affects the performance rather a combination of factors like socio-economic - cultural factors also have an impact. The data was filtered using manual techniques and saved as an artiff file. The open source tool WEKA was used for experimentation which contains many machine learning algorithms for data mining [17].Table 1 shows the statistical result of collected data set. Table 2 shows comparison of the classification algorithm for the collected data set. Table 3. Show the statistical analysis of downloaded data set. Table 4 shows the comparison of the classification algorithm for a downloaded data set.

5 RESULTS

Data set is collected from high school and downloaded from the report builder then tested using five classification algorithm those are Multilayer Perception, Naive Bayes, KStar, PART and J48. Results are provided in table 1 and table 2.

<table>
<thead>
<tr>
<th>Name of the classification algorithm</th>
<th>Class</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multilayer</td>
<td>P</td>
<td>1.000</td>
<td>0.273</td>
<td>0.961</td>
<td>1.000</td>
<td>0.980</td>
<td>0.997</td>
</tr>
<tr>
<td>Perceptron</td>
<td>F</td>
<td>0.727</td>
<td>0.000</td>
<td>1.000</td>
<td>0.727</td>
<td>0.642</td>
<td>0.904</td>
</tr>
<tr>
<td>Naive</td>
<td>P</td>
<td>0.932</td>
<td>0.182</td>
<td>0.971</td>
<td>0.932</td>
<td>0.951</td>
<td>0.997</td>
</tr>
<tr>
<td>Bayes</td>
<td>F</td>
<td>0.818</td>
<td>0.068</td>
<td>0.643</td>
<td>0.818</td>
<td>0.720</td>
<td>0.924</td>
</tr>
<tr>
<td>KStar</td>
<td>P</td>
<td>1.000</td>
<td>0.455</td>
<td>0.936</td>
<td>1.000</td>
<td>0.967</td>
<td>0.998</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0.545</td>
<td>0.000</td>
<td>1.000</td>
<td>0.545</td>
<td>0.706</td>
<td>0.962</td>
</tr>
<tr>
<td>PART</td>
<td>P</td>
<td>0.986</td>
<td>0.182</td>
<td>0.973</td>
<td>0.986</td>
<td>0.980</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0.818</td>
<td>0.014</td>
<td>0.900</td>
<td>0.818</td>
<td>0.857</td>
<td>0.771</td>
</tr>
<tr>
<td>J48</td>
<td>P</td>
<td>0.986</td>
<td>0.182</td>
<td>0.973</td>
<td>0.986</td>
<td>0.980</td>
<td>0.994</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>0.818</td>
<td>0.014</td>
<td>0.900</td>
<td>0.818</td>
<td>0.857</td>
<td>0.771</td>
</tr>
</tbody>
</table>
From the simulation results it was observed that the KStar is giving highest frequency in the downloaded data set and also it was observed that the PART and J48 is giving highest frequency in the collected data set. As an extension of this work the various analysis of relationship between the various attributes in finding the slow learners, the relationship between attendance and the student performance in various subjects etc will be decided. This will also be used for predicting the performance of each student well in advance so as to enable the authorities to take appropriate remedial actions to improve the overall performance.

6 CONCLUSION
From the simulation results it was observed that the KStar is giving highest frequency in the downloaded data set and also it was observed that the PART and J48 is giving highest frequency in the collected data set. As an extension of this work the various analysis of relationship between the various attributes in finding the slow learners, the relationship between attendance and the student performance in various subjects etc will be decided. This will also be used for predicting the performance of each student well in advance so as to enable the authorities to take appropriate remedial actions to improve the overall performance.

ACKNOWLEDGMENT
I would like to thank Prof. Deepa V. Jose my research guide and Sri Barathi Vidhyalaya hr.sec School, Hosur, Tamil nadu.

REFERENCES


Fig-1.Graph representation for comparison of classification algorithm on collected data set

Fig-2.Graph representation for comparison of classification algorithm on downloaded data set


