AUTOMATION OF PRE-PROCESSING OF STUDENTS DATA

Sridevi Bonthu

Department of Computer Science and Engineering, Vishnu Institute of Technology,
Bhimavaram, Andhra Pradesh, India
sridevi.db@gmail.com

B V Prasanthi

Department of Computer Science and Engineering, Vishnu Institute of Technology,
Bhimavaram, Andhra Pradesh, India
prasanthibeera@gmail.com

K Himabindu

Department of Computer Science and Engineering, Vishnu Institute of Technology,
Bhimavaram, Andhra Pradesh, India
himabindu.k@vishnu.edu.in

Abstract

The main challenge of any educational institution today is explosive growth of educational data and its usage in taking decisions in the improvement of quality decisions. Institutions follow their own templates to maintain the student data. Data need to undergo pre-processing before knowledge extraction. Pre-processing this data will become tedious, if one wants to gain knowledge on the data of more than one institution. Automation of pre-processing resolves this issue. This paper addresses how data of any organization can be pre-processed by using the tool developed by the authors. The main objective of this work is to minimize user involvement and maintain data integrity in the student data of educational institutions.

***Keywords*:** Data pre-processing; Automation; Educational Data Mining; Python.

1. Introduction

Data mining is a technique to analyze and extract knowledge from large amount of data which can identify the patterns and correlations that exist in the massive amount of data. The data mining process applies various algorithms to perform classification, clustering, association, prediction to acquire the knowledge for decision-making [1]. Educational Data Mining (EDM) is a field that exploits Data Mining (DM) algorithms in different types of educational data in order to resolve educational research issues [2]. Higher educational institutions are part of service industry and they treat students as their prime customers. In the current times institutions are following plenty of approaches to improve their student's performance, to improve quality in education, reduce student drop-out ratio, identify slow learners etc. These approaches can be decided by analyzing the existing student performance data. Government can take decisions to improve student community, by analyzing student performance from all the institutions in a region. Based on this architecture shown [3,4] any educational institution can build their own storage cloud to provide storage to this kind of data. Maintaining the data is one issue and Providing security [5] to that data is an another issue, this can be achieved by the combination of biometrics. Even there are tools [6] to provide the security.

Data mining play an important role in decision making, but inconsistencies and noise present in it may result in poor and erratic decisions. It is important to consider that the real world data is prone to inconsistencies, duplicate values and can use dissimilar units for same attribute. Mining with that kind of data may give unreliable knowledge model [7]. The results of any data mining algorithm completely depend upon the quality of the data. Data pre-processing consumes nearly 50% of the efforts for knowledge extraction. Though data pre-processing is one of the most important step, but less studied task in educational data mining research [8] clean and tidy data is suitable for application of algorithms leading to better knowledge.

Pre-processing of data manually can significantly delay downstream analysis and increase the possibility of human errors. Data is collected from various institutions to analyze. Different organizations follow different mechanisms to maintain data. Differences can be in the form of how they assign grades, maintain attendance, maintain past history, assignments etc. This paper focuses on automation of data pre-processing, which suits any

kind of educational institution. This tool works by asking as set of questions relevant to educational data and converts their data to a valid dataset.

In colleges the student performance is judged by analyzing their marks. It is easy to analyze data when the dataset is small, but in the case of large volumes of data it will be difficult to analyze. Analysis of student dataset is performed by using different attributes like attendance, internal marks, external marks, assignments, backlogs etc., before knowledge extraction, the dataset need to be pre-processed. This paper showcases an automated pre-processing tool to perform that initial step. Literature review of pre processing techniques is presented in section 2 followed by the architecture of the tool in section 3. Demonstration of usage of the tool on the student's data of Vishnu Institute of Technology is in section 4 and Conclusion in section 5.

2. Related Work

Data cleaning makes the data suitable as input for certain algorithms and leads to correct and usable knowledge. The data pre-processing phase typically requires a significant amount of manual work, which may take up 60–90 % of the time, efforts and resources employed in the whole knowledge discovery process [9].

This tool may become helpful for any kind of higher educational institution which wants to analyze their wards. It accepts any type of dataset either small or large to analyze. This tool is developed using Enthought Canopy (IDE) [10] and some Python libraries [11] like scikit-learn, numpy, pandas, matplotlib.

Educational systems provide a huge amount of student information generated every day from different sources of information [12]. It is humanly impossible to study, decipher, and interpret all that data to find useful information [13]. There are typically a great number of attributes available about students and a lot of instances at different levels of granularity. So, it is required to use attribute selection and filtering tasks in order to select the attributes and instances that can help to address a specific educational problem [12]. Data pre-processing manually consumes more time. Sometimes it may become more the than the time taken for knowledge discovery in datasets. The paper comes up with an automation tool after observing the data present at various institutions. Identification of the attributes which play major role indecision making is done to design this tool. The following design model is followed to automate the pre-processing.

3. Architecture

In order to facilitate the application to do automation, software architecture is adopted. Figure conveys that the designed tool can take data from any educational institution and convert into a dataset which is suitable to perform data analytics. Data is cleaned first by handling missing and noisy data. There are several techniques that can be used to deal with missing data. Choosing the right technique is an option that depends on the problem domain—the data's domain and our objective for the data mining process Unnecessary attributes which are not going to play any role in educational data mining were removed. Data transformation transforms or consolidates data into forms, which is appropriate for mining, by performing summary or aggregation operations. Data is not collected only for data mining. Therefore to downsize the data, dimensionality reduction is an effective approach.

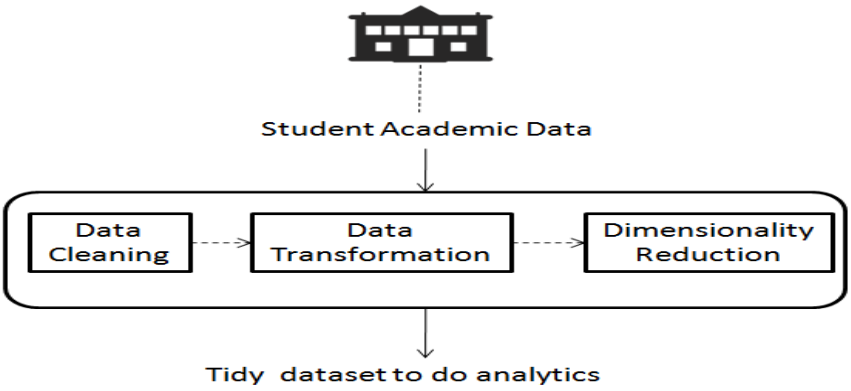

Fig. 1. Software Architecture of Pre-processing tool

4. Methodology

This new tool is designed based on the user (faculty, management) requirements and detailed analysis of the current system. In this work, the raw data will be converted into an efficient dataset by undergoing an automated

pre-processing tool. The techniques followed to automate pre-processing, filling missing values, conversion of continuous variables to discrete or categorical variables, normalizing a column, and encoding categorical features are furnished below.

The main attributes this tool excepts are Gender, Attendance, internal marks, external marks, assignment marks, backlogs etc. The column names are furnished below for a sample data frame.

```
['Regd no', 'Name', 'Gender', 'Attn', 'SSC', 'INTER', 'INT1_1',
 'EXT1_1', 'BK1_1', 'INT1_2', 'EXT1_2', 'BK1_2', 'Q1', 'Q2', 'Q3',
 'Q4', 'Q5', 'Q6', 'Qavg'], dtype=object)
```

The clear description of the attributes is furnished in the below table.

Table 1. Attributes of the dataset

Attributes	Remarks
Gender	Gender of the student. Can be M or F / Male or Female / 0 or 1. Plays an important role in some analytics problems.
Attendance	Specified in percentages. Ranges from 0 to 100
SSC, Inter	Percentage of SSC and Intermediate respectively. They range from 0 to 100. These are measures to evaluate students' past record
Internal Marks	Number of midterm examination conducted between 1–2 among all the semesters taken into consideration.
External Marks	Marks of University Examinations. These important attributes to assess the students
Backlogs	Contains number of subjects failed semester wise.
Quiz marks	six quizzes are counted as per the course policy which was intact throughout all observed semesters.

These attributes help to analyze the student performance. The sample dataset is taken in the form of a csv file. The tool even takes an excel file also as input source. This tool provides a user interface through which one can open the data file and upload. Afterwards it converts it into a valid data frame of python pandas. The sample dataset shown in this paper is real world data and is collected from an engineering college (Vishnu Institute of Technology, Bhimavaram).

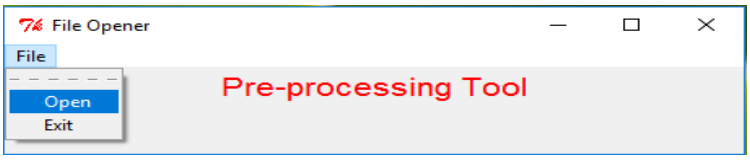

Fig. 2. GUI of the pre-processing Tool

This pre-processing tool works by asking a set of questions regarding the pre-processing techniques.

- This tools first step is to handle missing values. Tool displays all its attributes and allows the user to select some to fill those missing values. System provides few options to the college management for filling the missing values. We can use mean or mode of attribute values or can even use some user-specified knowledge to fill the missing values.
- Tool takes some information while translating few fields, as this is designed to be able to work with any educational institution. The attendance attribute in the student dataset may be in the form of percentages. It is better if the percentage is converted into ranges to improve the accuracy of prediction models and data analytics. Hence, the tool allows the user to select the name of attendance attribute, as different institutes may maintain different attribute names and transforms them into ranges.
- The raw dataset may contain attributes with different values. In order to scale them into certain ranges the tool allows the user to select attributes and uses some methods that are available in python language to scale them.
- The gender attribute usually contains two values MALE and FEMALE. But some clustering algorithms like k-means don't support string, so we need to convert them into 0 and 1. This tool converts the gender column into 0 and 1.

5. Results

The primary data required is taken from Vishnu Educational Institutions, Andhra Pradesh. Student data with various attributes and varying dimensions are fed to this designed tool. Following figure shows the top 10 rows

of the sample data set. The Gender is specified as M and F, and the attendance is marked in percentages and some missing values were present in the quiz marks.

Regd no	Name	Gender	Attn	SSC	INTER	INT1_1	EXT1_1	BK1_1	INT1_2	EXT1_2	BK1_2	Q1	Q2	Q3	Q4	Q5	Q6	Qavg
14PA1A0501	AADI BHARGAV SAI MAHESH KUMAR	M	91.38	68	71.0	137	138	5	139	169	6	5	5	0	0	4	5	3.166667
14PA1A0502	AKULA MADHAVI	F	94.83	93	94.9	222	409	0	224	471	0	5	4	3	NaN	NaN	NaN	4.000000
14PA1A0503	AKULA SAI PAVAN HEMANTH	M	62.07	80	86.0	156	207	4	158	241	3	3	5	0	1	0	5	2.333333
14PA1A0504	AKULA SIVA NAGA VENKATA SRI RAM	M	81.03	88	90.2	221	386	0	228	427	0	5	5	5	NaN	NaN	NaN	5.000000
14PA1A0505	ALAGINGI HEMAMALINI	F	72.41	93	90.4	220	428	0	228	483	0	5	5	5	NaN	NaN	NaN	5.000000
14PA1A0506	ALAPATI JAHNAVI	F	67.24	93	92.2	210	346	0	212	468	0	5	5	4	NaN	NaN	NaN	4.666667
14PA1A0507	ALLURI SAI NIKHIL VARMA	M	86.21	72	78.0	185	270	1	187	302	1	4	4	2	3	4	NaN	3.400000
14PA1A0508	ANNAMBHOTLA V NARASIMHA SITA SOWMYA SREE	F	25.86	75	80.0	188	304	0	199	313	2	3	3	3	5	5	9	4.666667
14PA1A0509	ARIGE SYAMALA DEVI	F	75.86	77	87.0	205	340	0	201	440	0	5	4	4	NaN	NaN	NaN	4.333333
14PA1A0510	ATTILI NAGA VARA SAI SOMA SUNDAR	M	43.10	93	63.0	114	173	4	117	152	5	0	0	0	NaN	NaN	NaN	0.000000

Fig. 3. The dataset before pre-processing

The figure shows the resultant dataset given by the tool. If we observe the results, gender attribute is translated into 0 and 1. Attendance is replaced by some numbers in place of percentages. SSC and Inter marks are past history of the student. Only SSC column is scaled by the tool to show the difference. Missing values of quiz fields were filled.

	Name	Gender	Attn	SSC	INTER	INT1_1	EXT1_1	BK1_1	INT1_2	EXT1_2	BK1_2	Q1	Q2	Q3	Q4	Q5	Q6	Qavg
Regd no																		
14PA1A0501	AADI BHARGAV SAI MAHESH KUMAR	1	3	0.265823	71.0	137	138	5	139	169	6	5	5	0	0.0	4.000000	5.000000	3.166667
14PA1A0502	AKULA MADHAVI	0	3	0.898734	94.9	222	409	0	224	471	0	5	4	3	3.8	4.275862	4.964286	4.000000
14PA1A0503	AKULA SAI PAVAN HEMANTH	1	2	0.569620	86.0	156	207	4	158	241	3	3	5	0	1.0	0.000000	5.000000	2.333333
14PA1A0504	AKULA SIVA NAGA VENKATA SRI RAM	1	3	0.772152	90.2	221	386	0	228	427	0	5	5	5	3.8	4.275862	4.964286	5.000000
14PA1A0505	ALAGINGI HEMAMALINI	0	2	0.898734	90.4	220	428	0	228	483	0	5	5	5	3.8	4.275862	4.964286	5.000000
14PA1A0506	ALAPATI JAHNAVI	0	2	0.898734	92.2	210	346	0	212	468	0	5	5	4	3.8	4.275862	4.964286	4.666667
14PA1A0507	ALLURI SAI NIKHIL VARMA	1	3	0.367089	78.0	185	270	1	187	302	1	4	4	2	3.0	4.000000	4.964286	3.400000
14PA1A0508	ANNAMBHOTLA V NARASIMHA SITA SOWMYA SREE	0	0	0.443038	80.0	188	304	0	199	313	2	3	3	3	5.0	5.000000	9.000000	4.666667
14PA1A0509	ARIGE SYAMALA DEVI	0	2	0.493671	87.0	205	340	0	201	440	0	5	4	4	3.8	4.275862	4.964286	4.333333
14PA1A0510	ATTILI NAGA VARA SAI SOMA SUNDAR	1	1	0.898734	63.0	114	173	4	117	152	5	0	0	0	3.8	4.275862	4.964286	0.000000

Fig. 4. The dataset after pre-processing

6. Conclusion

Pre-processing allows transforming the available raw educational data into a suitable format ready to be used by a data mining algorithm for solving a specific educational problem. With the tool shown in paper, pre-processing of education data is automated. With this tool user involvement is minimized in the student data of educational institutions.

Acknowledgments

The authors would like to thank the Department of Computer Science and Engineering, Vishnu Institute of Technology for providing the students' course data for testing the pre-processing tool.

References

[1] Christa, S., Madhuri, L., & Suma, V. (2012). An Effective Data Preprocessing Technique for Improved Data Management in a Distributed Environment. In International Conference on Advanced Computing and Communication Technologies for High Performance Applications, International Journal of Computer Applications, Cochin (pp. 52-57).

[2] Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12-27.

[3] Bonthu, S., & Srilakshmi, M. (2014). Building an Object Cloud Storage Service System using OpenStack Swift. International Journal of Computer Applications, 102(10).

[4] Bonthu, S., Murthy, Y. S. S. R., & Lakshmi, M. S. (2014). An Adaptive Downloading Service from Object Storage Cloud. International Journal, 4(8).

[5] Prasanthi, B. V. "Security Enhancement of ATM System with Fingerprint and DNA Data." International Journal of Advanced Research in Computer Science and Software Engineering 4.12 (2014): 477-479.

[6] Prasanthi, B. V. "Cyber Forensic Tools: A Review." International Journal of Engineering Trends and Technology (IJETT) 41.Number-5 (2016): 6.

[7] Christa, S., Madhuri, L., & Suma, V. (2012). An Effective Data Preprocessing Technique for Improved Data Management in a Distributed Environment. In International Conference on Advanced Computing and Communication Technologies for High Performance Applications, International Journal of Computer Applications, Cochin (pp. 52-57).

[8] Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In Educational data mining (pp. 29-64). Springer International Publishing

[9] Gonçalves Jr, P. M., Barros, R. S., & Vieira, D. C. (2012, May). On the use of data mining tools for data preparation in classification problems. In Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on (pp. 173-178). IEEE.
[10] Python Distribution and Integrated Analysis Environment | Enthought Canopy. (2017). Enthought.com. Retrieved 27 January 2017, from https://www.enthought.com/products/canopy/
[11] PyPI - the Python Package Index. (2017). Pypi.python.org. Retrieved 27 January 2017, from https://pypi.python.org
[12] Sael, N., Marzak, A., & Behja, H. (2012). Investigating an advanced approach to data preprocessing in moodle platform. International Review on Computers and Software, 7(3).
[13] Thakur, R., & Mahajan, A. R. (2015). Preprocessing and Classification of Data Analysis in Institutional System using Weka. International Journal of Computer Applications, 112(6).