ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

Effect of normalization as a preprocessing task in result declaration

Nilesh Mahajan, Jyoti Namdeo
Professor, Research Scholar

Abstract*— Educational Data Mining is concerned with developing methods for exploring the unique types of data that come from educational settings to better understand students. The choice-based credit system (CBCS) calls for a fair and transparent internal assessment. Credit based system allows the students to study subjects of their choice along with the core subjects of their faculty at post-graduate and doctorate level. Present work is aimed to understand the effect of normalization on the results of MCA students from traditional system into the credit based system and to apply different data mining tools like discretization and correlation analysis. The original data is first converted to grade point for CBCS pattern and the subjects under consideration are correlated. This data is compared with normalization of original data to grade point system followed by correlation analysis. The correlation analysis of the subjects under consideration showed that discretization after normalization provides better result.

***Index Terms*— Correlation analysis, choice-based credit system (CBCS), discretization, educational data mining, normalization.**

I. INTRODUCTION

Educational data mining is to learn the educational data and show up the hidden knowledge from it. The huge databases contain a wealth of data and constitute a potential goldmine of valuable information. Educational organizations consider students and teachers as their main assets and want to improve their key process indicators by effective and efficient use of their assets.

In India, for the academic reforms of higher educational system, a uniform academic calendar, introduction of Choice based Credit System (CBCS) and semester system, examination reforms including continuous internal assessment and grading system are recommended by academic commissions and committees such as UGC and NAAC.

Choice based system (CBCS), or a cafeteria like system is the solution for transformation from the traditional teacher oriented education to a student-centered education. In this way students can take responsibility for their own education and can be benefitted most from all the available resources. CBCS has several unique features such as enhanced learning opportunities, ability to match student's scholastic needs and aspirations, inter institution transferability of students, part completion of an academic program in the institution of enrollment and part completion in a specialized (and recognized) institution, improvement in educational quality and excellence, flexibility for working students to complete programme over an extended time and standardization and compatibility of educational programmes across the country.

II. REVIEW OF LITERATURE

Extracting hidden information from a huge set of data is an important and a challenging task in data mining. Data with credibility and relevance plays a vital role and it is important to ensure that genuine and good quality data is being used (Jansi Rani and Bhaskaran 2010).

Data are normally preprocessed through data cleaning, data integration, data selection, data transformation and prepared for the mining task. Advancing statistical methods and machine learning techniques have played important roles in analyzing high dimensional data sets for discovering patterns hidden in it Dash et. al. (2010).

Vialardi et. al. (2009) developed recommendation system based on data mining techniques to help students to take decisions on their academic itineraries, to choose course, based on experience of previous students with similar academic achievements. However, Sahay and Mehta (2010) developed a software system to assist higher education in assessing and predicting key issues related to student success. Software used data mining algorithm

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

and quality tools such as quality function deployment to study and predict issues related to enrollment management, dropout rate, and time to degree and suggest ways to improve courses and programs.

Hsia et. al. (2008) analyzed the course preferences and course completion rates of enrollees in extension education courses at a university in Taiwan. They analyzed data by using three data mining algorithms: Decision Tree, Link Analysis, and Decision Forest.

Ayesha et. al. (2010) studied K-means clustering algorithms to discover knowledge from education data mining. They recommended that all correlated information of class quiz, mid and final exam should be conveyed so that dropout ratio can be reduced and student performance can be improved. Quadril and Kalyankar (2010) used data mining in predicting drop out feature of students. They used decision tree techniques to choose the best prediction and analysis for direct or indirect intervention from teacher and management.

Namdeo V. et. al. (2010) collected student's data from engineering and applied four different classification methods and classifies students based on their final grades and applied four classification methods on student data i.e. Decision tree (ID3), Multilayer Perceptron, Decision Table and Naïve Bayes Network Classification method. They concluded that ID3 Classifier is most suitable method for this type of student dataset. The same result was also proved by Cristina Oprea and Marian Zaharia (2011) whose results indicated that the ID3 algorithm and Random Forest algorithm provided the highest accuracy and correctly classified instances. Mohamad Farhan et. al. (2010) focused the relationship between academic factor and personality characteristic towards programming performance.

There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data transformations, such as normalization, may be applied. Normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Normalization methods represent compromises designed to achieve particular ends. Normalization requires taking values that span one range and representing them in another range. This requires remapping values from an input range to an output range. Each method of remapping may introduce various distortions or biases into the data. Some biases and distortions are deliberately introduced to better expose information content. Others are unknowingly or accidentally introduced, and damage information exposure. Some types of bias and distortion introduced in some normalization processes are beneficial only for particular types of data, or for particular modeling methods Payle (1990), Karthikeyani et. al. (2009), Frankes and Baeza-Yates (1992) Han and Kamber 2006.

Patel and Mehta (2011) compared three normalization techniques and improved the efficiency and quality of results. Similarly, Jain and Bhandare (2011) proposed a privacy preserving data distortion method based on min max normalization transformation. Saranya and Manikandan (2013) also analyzed the use of normalization techniques to preserve data privacy. In another study, Syed et. al. (2014) dealt with min- max normalization based data transformation to preserve data privacy.

III. DISCRETIZATION OF DATA

The secondary data is obtained from the examination section of a University in Pune, India. Masters in Computer Applications (MCA) is a professional program of three years having six semesters. Each semester has 7 subjects and thus the total number of subjects including compulsory and elective is about 40. The first step of data mining is discretization of data. On the basis of literature survey, it was revealed that discrete values plays an important role in data mining and knowledge discovery because it deals with intervals of numbers which are more concise to represent and specify. Discrete values are easier to use and comprehend as they are closer to a knowledge-level representation than continuous values.

Discretization of real value attributes (features) is an important pre-processing task in data mining. Various techniques have been studied for converting the continuous data into discrete one. Chmielewski and Grzymala-Busse, 1994; Dougherty et al. 1995; Nguyen and Skowron, 1995; Nguyen, 1998; Liu et al., 2002. Dougherty et al., 1995 have suggested that discretization makes learning faster.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

For the present study, the conditional attributes are the compulsory theoretical subjects in MCA program, whose performance decide the result of students. The numbers of such conditional attributes are about 20. Marks obtained by students in these compulsory subjects ranges from 0-100. Instead of dealing with the whole range of marks, the marks scored by students are subdivided into groups for better understandability.

University gives the marks out of hundred. These hundred marks are in the ratio of 80:20. University conducts the main exam (i.e. theoretical out of 80), whereas 20 marks form the basis of internal assessment. In the present study we concentrated only on the main exam conducted by the university leaving the internal marks as we are interested only in the study of performance of students in the theoretical exam. Since the marks are out of 80, we converted the obtained marks of student into percentage. In this paper we have used two ways of discretization.

A. The result of the student depends on the conditional attributes. The final result of student is declared by evaluating the average marks scored at the completion of the program. These marks contain only the total of theoretical marks out of 80.The 10-point scale is easy to handle the selected data as this discretize the continuous range of 0-100 marks into 9 discrete intervals.

Table 1: Grading system to discretize the range of marks into groups as 10 Grade Point.

MARKS	GRADE POINT
[75-100]	10
[70-74.9]	9
[65-69.9]	8
[60-64.9]	6
[55-59.9]	7
[50-54.9]	5.5
[45-49.9]	5.0
[40-44.9]	4.5
[00-39.9]	0

The final result of the students is also discretized. Table 2 illustrates the discretization of decision attribute (final results).

Table 2: Discretization of Decision Attribute (Final results)

MARKS	GRADES	EXPLAINATION
[70-100]	O	D, Distinction
[60-69.99]	A+	F, First class
[55-59.99]	B+	HS, Higher Second class
[50-54.99]	B	S, Second class
[40-49.99]	C	P, Pass class
[00-39.99]	F	Fl, Fail class

Using the above criteria, all the subject marks along with the result are discretized.

B. In the exam conducted by university, student gets marks. These marks are the raw marks. In the first method of discretization we converted the marks into percentage and then discretized them according to the corresponding grades. However these raw marks have some variations such as the effect of teacher who taught the subject, effect of paper pattern, effect of examiner's evaluation on the students answer sheet, effect of book that student referred etc. These variations can be removed by normalization of the raw data. This practice is not uncommon as all credit based system uses such normalized data for awarding grades to students. After calculating the average marks and standard deviation of individual subject, we computed the normalized data by using the following formula:

Normalized data $x' = \frac{x}{}$

Where x'= normalized mark
x = raw mark

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

μ = mean of the subject
σ = standard deviation of the subject

All the raw data is converted into the normalized data and hence different approach of discretization is used. In this method, for the present study subject mean and subject standard deviation both are used in making the discrete intervals which are represented in table 3.

Table 3: Intervals based on discretization

Interval	Grade
[min{32, (μ - σ)}, (μ-σ))	F
[(μ-σ), (μ - 0.5σ))	E
[(μ - 0.5σ), μ)	D
[μ, (μ + 0.5σ))	C
[(μ + 0.5σ), (μ + σ))	B
[(μ + σ), max{64, min{(μ + 1.5σ), 72}})	A
[max{64, min{(μ + 1.5σ), 72}}, 80]	O

Using the above criteria, all the subject marks along with the result are discretized.

IV. CORRELATION ANALYSIS

Correlation analysis is a statistical approach that finds relationships among pairs of variables. Such variables usually represent properties of objects whose values may be stored in columns of a database table. Correlations among variables can be negative or positive. There are different ways of computing correlations, but in most cases, the correlation is measured as a coefficient ranging from -1 to 1. A value close to 0 in this range indicates a lack of correlation. Values closer to the boundaries -1 or 1 indicate strong negative or positive correlations, respectively. Usually strong positive or negative correlations may indicate a causal relationship between the variables. For example, there may be a positive correlation between the number of hours of studying for an exam and the score obtained in that exam.

Karl Pearson's coefficient of correlation (or simple correlation) is the most widely used method of measuring the degree of relationship between two variables.
Karl Pearson's coefficient of correlation can be worked out thus.

Karl Pearson's coefficient of correlation (or r)* = $\sum(X_i - \bar{X})(Y_i - \bar{Y})/(n.\sigma_X.\sigma_Y$

Where r = Karl Pearson's coefficient of correlation
X_i = ith value of *X* variable
$\bar{X}$ = mean of *X*
Y_i = ith value of *Y* variable
$\bar{Y}$ = mean of *Y*
n = number of pairs of observations of *X* and *Y*
σ_X = Standard deviation of *X*
σ_Y = Standard deviation of *Y*

We were interested to find the relationship between any two subjects of the MCA course. So in order to find the association between subjects we concentrated only on the main 12 subjects. With the help of excel sheet we calculated the correlation coefficient between the pair of subjects. This calculation was performed on the absolute marks of the students. The corresponding correlation coefficients are shown in the following table3.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

Table 4. Correlation table of absolute data

	UE101	UE103	UE 201	UE 202	UE 203	UE 301	UE 302	UE 303	UE 401	UE 402	UE 501	UE 502
UE101	1	0.614544	0.421264	0.404834	0.325869	0.536146	0.370759	0.433089	0.445815	0.371704	0.357245	0.339377
UE103		1	0.581082	0.573706	0.506538	0.536146	0.442867	0.712891	0.602276	0.511303	0.501445	0.595103
UE201			1	0.794473	0.675322	0.706672	0.745496	0.544486	0.791804	0.616501	0.585198	0.457328
UE202				1	0.681642	0.728295	0.604792	0.609589	0.647339	0.660142	0.626441	0.408621
UE203					1	0.580065	0.57547	0.449399	0.640677	0.597399	0.562885	0.391764
UE301						1	0.644034	0.569453	0.759721	0.656981	0.582992	0.56202
UE302							1	0.436132	0.668115	0.677998	0.570836	0.352812
UE303								1	0.584363	0.685966	0.648426	0.593657
UE401									1	0.717568	0.683708	0.457667
UE402										1	0.627448	0.41114
UE501											1	0.424405
UE502												1

The same type of correlation coefficient was calculated for the normalized score using the same set of subjects. The correlation analysis was performed to find how strongly or weakly the two subjects are associated with each other. The correlation table for the normalized score is shown below. The value of r between any two variables i.e. subjects is coming comparatively low as that of value of r obtained after normalizing the data.

Table 5. Correlation table of normalized data

	N101	N103	N201	N202	N203	N301	N302	N303	N401	N402	N501	N502
N101	1.000	0.742013	0.609085	0.602158	0.522206	0.716824	0.623128	0.624168	0.623615	0.635268	0.600315	0.634789
N103		1	0.723386	0.726261	0.656032	0.700833	0.677231	0.80362	0.728939	0.712359	0.688323	0.721957
N201			1	0.863038	0.765813	0.800762	0.835615	0.689655	0.854396	0.758352	0.728917	0.66191
N202				1	0.778638	0.816707	0.771021	0.731567	0.755629	0.786465	0.762531	0.662974
N203					1	0.718652	0.718536	0.611132	0.739352	0.709357	0.716496	0.619833
N301						1	0.82094	0.744349	0.834926	0.833701	0.794984	0.83121
N302							1	0.659622	0.781121	0.84765	0.772908	0.747159
N303								1	0.719768	0.799875	0.779351	0.732554
N401									1	0.805607	0.789857	0.663038
N402										1	0.794476	0.784579
N501											1	0.758675
N502												1

V. RESULTS

It can be clearly seen that correlation factor for the absolute data is coming very less compared to the correlation factor of normalized data. This is due to the fact that all type of variations has been removed from the absolute data. From the table we can find that 201 subject i.e. data structure is strongly associated with 202 i.e. operating system, 301 i.e. software engineering and 401 i.e. UML.

It can be interpreted as if the student thoroughly understands the data structure than he can easily understand operating system, software engineering and UML.

103 i.e. procedure oriented programming is strongly with 303 i.e. object oriented programming.

202 i.e. operating system is strongly associated with 301 i.e. software engineering.

301 i.e. software engineering is associated with 302 i.e. computer networking, 401 i.e. UML, 402 i.e., UNIX and 502 i.e. Artificial intelligence.

401 i.e. UML is strongly associated with 402 i.e. UNIX.

One observation is that each successive semester subject is associated or dependent on the previous semester subject.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

VI. CONCLUSION

Mining the student marks is very useful in predicting the future trends. In the present work we have used the two different ways of discretizing the student subject marks and student final result. The result clearly shows that normalized data is better than the absolute data because the normalized data does not contain the effects of hidden variables.

.

REFERENCES

[1] S. Ayesha and T. Mustafa, "Data Mining Model for Higher Education System." European Journal of Scientific Research 43, 1 pp.24-29, 2010

[2] M. Behrouz et.al.), "Predicting student performance: an application of data Mining methods with the educational web-based system Lon-capa". 33rd ASEE/IEEE Frontiers in Education Conference. Boulder, CO. pp1-6, 2003.

[3] C. Chaudhary (2012). Assessment of adoption of Choice Based Credit System by India Universities, International Journal of Behavioral Social and Movement Sciences. Vol.01, Issue 02. April 2012.

[4] Chmielewski and B Grzymala,"Global discretization of continuous attributes as preprocessing for machine learning". In Third International Workshop on Rough Sets and Soft Computing, pp. 294–301, 1994.

[5] R. Dash, R. Dash and D. Mishra, "A Hybridized Rough-PCA Approach of Attribute Reduction for High Dimensional Data Set", European Journal of Scientific Research Vol.44 No.1, pp.29-38, 2010.

[6] P. Dorian, "Data Preparation for Data Mining", Morgan Kaufmann Publishers, Inc. USA, p-225-227. 1990.

[7] Dougherty et al.,). Supervised and unsupervised discretization of continuous features. In Proc. Twelfth International Conference on Machine Learning. Los Altos, CA: Morgan Kaufmann, pp. 194– 202, 1995.

[8] M. Farhan et.al., "An Investigation into Influence Factor of Student Programming Grade Using Association Rule Mining." Advances in Information Sciences and Service Sciences Volume 2, Number 2, 2010.

[9] W. Frankes and R. Baeza-Yates,."Information Retrieval: Data Structures and Algorithms", Prentice–Hall, Englewood cliffs, NJ, 1992.

[10] Government of Gujarat," Educational Department, Implementation of Choice Based Credit System," No. CBC-262011-918-KH, Sachivalaya, Gandhinagar, dtd 11th April 2011.

[11] J. Han, and M. .Kamber, "Data Mining: Concepts and Techniques", Second edition, Morgan Kaufmann, USA, 2006.

[12] J. Han, , X. Hu and T. Y. Lin, "Feature Subset Selection Based on Relative Dependency between Attributes. Rough Sets and Current Trends in Computing": 4th International Conference, RSCTC 2004, Uppsala, Sweden, , pp. 176–185, June 1-5, 2004.

[13] T. Hsia et. al. Course planning of extension education to meet market demand by using data mining techniques – an example of Chinkuo Technology University in Taiwan. Expert Systems with Applications, vol. 34, pp. 596–602, 2008.

[14] Y. K Jain. and S. K., Bhandare, "Min Max Normalization Based Data Perturbation Method for Privacy Protection", International Journal of Computer & Communication Technology Volume-2 Issue-VIII, pp. 45-50, 2011.

[15] P. G. Jansi Rani. and R. Bhaskaran, "Extraction of Dominant Attributes and Guidance Rules for Scholastic Achievement Using Rough Set Theory in Data Mining," IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 10, pp. 28-37, May 2010,.

[16] N. Karthikeyani Visalakshi.and K. Thangavel, "Impact of Normalization in Distributed K-Means Clustering", International Journal of Soft Computing 4, Vol. 4, pp. 168-172, 2009.

[17] M .Lee, "Mining students' behavior in web-based learning programs." Expert Systems with Applications 36, pp. 3459–3464, 2009.

[18] H. Liu et al., "Discretization: An Enabling Technique," Data Mining and Knowledge Discovery, Kluwer Academic Publishers, the Netherlands, vol.6, pp.393–423, 2002.

[19] V. Namdeo et. al.," Result Analysis Using Classification Techniques." International Journal of Computer Applications. 1 – No. 22.pp. 0975 – 8887, 2010.

[20] H.S..Nguyen. and Skowron, "Quantization of real value attributes: rough set and Boolean reasoning approach". Proceedings of the Second Joint Annual Conference on Information Sciences, Wrightsville Beach, NC, pp. 34-37, 1995

[21] H.S. Nguyen, " Discretization problem for rough sets methods", Proceedings of the First International Conference on Rough Sets and Current Trends in Computing, Springer-Verlag., Warsaw, Poland, pp. 545-552, June 1998,

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 4, Issue 2, March 2015

[22] C.Oprea and M. Zaharia, "Using data mining methods in knowledge management in educational field". Fascicle of Management and Technological Engineering, Volume X (XX), NR1, 2011.

[23] V. R. Patel and R. G. Mehta," Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, and No 2. 2011.

[24] M. Quadri and N. Kalyankar," Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques", Journal of Computer Science and Technology 2 Vol. 10 Issue 2 (Ver 1.0), 2010

[25] A. Sahay and K. Mehta," Assisting Higher Education in Assessing, Predicting, and Managing Issues Related to Student Success: A Web-based Software using Data Mining and Quality Function Deployment". Academic and Business Research Institute Conference, Las Vegas. 2010.

[26] C. Saranya, and G. Manikandan," A Study on Normalization Techniques for Privacy Preserving Data Mining. International Journal of Engineering and Technology (IJET), vol.5, No.3 Jun-Jul, pp. 2701-2704, 2013.

[27] Md. Syed, ATarique, H. Shameemul and S. Prince Khan,. "Privacy Preserving in Data Mining by Normalization." International Journal of Computer Applications Volume 96– No.6, pp. 0975 – 8887, 2014

[28] University of Mumbai, "On Semester Based, Credit and Grading System For Under Graduates (UG) Programmes Under The Faculty of Commerce With Effect from the Academic Year 2011-12 COE_EXAM Approved by A. C. & M. C. ",_Manual_Commerce_June 2011.

[29] C.Veilardi et.al.," Recommendation in Higher Education Using Data Mining Techniques. "Educational Data Mining, pp. 190-199, 2009

AUTHOR BIOGRAPHY

Dr. Nilesh Mahajan: Dr. Mahajan is Professor in Institute of Management and Enterpreanurship development (IMED), Bharati Vidyapeeth Deemed University Pune, India

Mrs. Jyoti Namdeo: She is postgraduate in Computer Applications (MCA) and research scholar in Institute of Management and Enterpreanurship development (IMED), Bharati Vidyapeeth Deemed University Pune, India