IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 2, February 2017

Fuzzy Rule Based Classifier for Student Data using Map Reduce

Gayatri Nair[1], Shankar M. Patil[2]

Student, Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai, India[1]

Associate Professor, Department of Information Technology, Bharati Vidyapeeth College of Engineering,

Navi Mumbai, India[2]

Abstract: In the world of technology, enormous and complex data i.e. termed as Bigdata can be easily handled using latest processing tools and methods. One of the field dealing with such vast data is the education system for which various applications for processing, classifying and so on has been implemented. Various works include classification of student data to evaluate the performance of the student, Mining Social Media Data for Understanding Students' Learning Experiences, evaluating and predicting student performance before admission to the college as well as evaluating the suitability of the entry exams. Taking reference of these works classification of students academic dataset inorder to find which university he/she can get admission into is the focus of this paper. To be able to deal with this problem definition the usage of fuzzy rule based classification system has been proposed, using Fuzzy ID3 algorithm. This method is based on the MapReduce framework, one of the most popular approaches for big data nowadays.

Keywords: Big data, Fuzzy rule, Classification, Hadoop, Mapreduce, uncertainty.

I. INTRODUCTION

For the incredible increment in the data generation obtaining effective models that are able to conduct predictive analysis and extract knowledge from these huge data sources is necessary. It is a tedious job for users to find accurate data from huge unstructured data. So, there should be some mechanism which classifies unstructured data into organized form which helps user to easily access required data[8].As mentioned earlier, predictive analysis is an approach to deal with such enormous data. Predictive analytics is "an area of statistical analysis that deals with extracting information using various technologies to uncover relationships and patterns within large volumes of data that can be used to predict behaviour and events."[9][10].

Fuzzy logic first proposed by Zadeh in 1965.The concept of fuzzy logic founded to help computers deal with uncertain and ambiguous information [4]. Classification techniques over big transactional database provide required data to the users from large datasets more simple way. Classification technique is used to solve the challenges (include analysis, capture, datacuration, search, sharing, storage,transfer, visualization, querying and infor mation privacy) which classify the big data according to the format of the data that must be processed, the type of analysis to be applied, the processing techniques at work, and the data sources for the data that the target system is required to acquire, load, process, analyze and store.[2] Before actual classification begins, required information is extracted from large amount of data and then classification is done. The fuzzy logic is used under any one of the 2 conditions i.e. If the problem definition is not clear or ambiguous or, when the application is clear but the solution is vague[1]. The frameworks that are typically used to handle big data somehow involve some kind of parallelization so that they can easily process and analyse the data that is ready to be used.

II. RELATED WORK

Some of the earlier related work include the development of fuzzy systems using different methods and algorithms. One of them, efficient way to classify student grades, which represented by student level of knowledge according to multiple criteria. For example, student test degree and test time along with test level of complexity of the online-test. The proposed model designed and tested to evaluate student grades in Java Programming language. The output results showed that using fuzzy rule base system with multiple conditions improve grade classification process in online test systems. They used Mamdani technique[6]. Analytics is the process of discovering, analyzing, and interpreting meaningful patterns from large amounts of data. Following this statement and other research works, the author has tried to convey how predictive analysis has been used at a variety of institutions, including a review of its potential pitfalls and benefits. Also, has recommended to all colleges and universities to consider building predictive analytics into their toolbox of techniques that inform and enable evidence-based decision-making [9]. A Neuro-Fuzzy Classification Approach to the assessment of student performance used the concept of neural network and fuzzy

DOI 10.17148/IJARCCE.2017.6216

logic. The application was evaluate and predict student performance before admission to the college as well as evaluating the suitability of the entry exams. The resulting model would be used to support the student admittance procedure[1]. Using linguistic Fuzzy rules the classification system was developed in Map-reduce framework. They called the methodology as Chi-FRBCS-BigData and built 2 versions of it as Chi-FRBCSBigData-Max and Chi-FRBCS-BigData-Ave. The results show that the proposal is able to provide competitive results, obtaining more precise but slower models in the Chi-FRBCS-BigData-Ave alternative and faster but less accurate classification results for Chi-FRBCS-BigData-Max[8].

III. METHODOLOGY

To be able to deal with big data the usage of a fuzzy rule based classification system has been proposed using ID3 algorithm. As a fuzzy method, it is able deal with the uncertainty that is inherent to the variety and veracity of big data[8]. The proposed system looks into the academic scores of the students and classifies the students on the basis of their scores to various Universities. This method is based on the MapReduce framework, one of the most popular approaches for big data nowadays. Fuzzy Rule Based Classification Systems (FRBCSs) are potent and popular tools for pattern recognition and classification[8]. In my work, am considering the student data set which includes attributes as unique id, name, each section marks, total and pass/fail status. Decision tree is generated using Fuzzy ID3 algorithm. ID3 has highly unstable classifiers with respect to minor perturbation in training data. Fuzzy logic brings in an improvement of these aspects due to the elasticity of fuzzy sets formalism. Thus, ID3 was further modified into Fuzzy ID3-combination of fuzzy and mathematics [7]. By comparing the university cut-off and student cut-off, list of universities are displayed. Some students having less cut-off are still listed; reason being their student section wise marks are fitted in the University section wise marks. And, for those who doesn't fit in any of these criteria are displayed with Sorry message.

IV. IMPLEMENTATION

The classification in the proposed system is being done using the decision tree. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

The core algorithm for building decision trees called ID3 by J. R. Quinlan. ID3 uses Entropy and Information Gain to construct a decision tree. Fuzzy logic brings in an improvement of these aspects due to the elasticity of fuzzy sets. Fuzzy ID3 is only an extension of the ID3 algorithm achieved by applying fuzzy sets.In fuzzy ID3 the dataset is considered to be continuous [7]. Hence membership values for each of the attribute is defined. Thereafter the entropy and information gain is calculated.

A. Fuzzy ID3 algorithm:

1. Calculate entropy and Information Gain for each attribute i.e the average columns and pass/fail column.

$$Hr(S,A)= \sum_{i}^{c}{}_{=1}\sum_{j}^{n} \mu ij / S \ \log_2 \sum_{j}^{n}\mu ij / S$$

$$Gr(S,A) = Hr(S,A) - \sum NVEA |Sv|/S *Hfs(S,A)$$

Where,

μij= membership value of the jth pattern to ith class

|Sv|= size of subset S, Sv is the subset of training samples xj with ‘v’ attributes.

Hf(s) = entropy of set S of training samples in node

A membership function (MF) is a curve that defines how each point in the input space is mapped to a membership value between 0 and 1.Here,membership value for each attribute is defined. For example avg_term_mrks can be within the range of 20. So, all the scores will be mapped in graphical format from 0 - 20 as the MF 0,1.

2. Calculate the highest gain

Avg_exam_sec has given the highest gain. Hence it'll be considered as the root node

3. Define some threshold(cut-off) to prune the free on minimize the rules Fuzzy control threshold | if ratio of class ck > this threshold stop expanding tree Leaf decisions | if no data is > this threshold then stop expanding tree

4. Generate root node with membership value & generate sub nodes whose membership value is product of original membership value. Avg_exam_sec is the root node and it branches into the range of scores i.e 0 to 35 and 35 to 60. Similarly further branches and finally leaf node is generated.

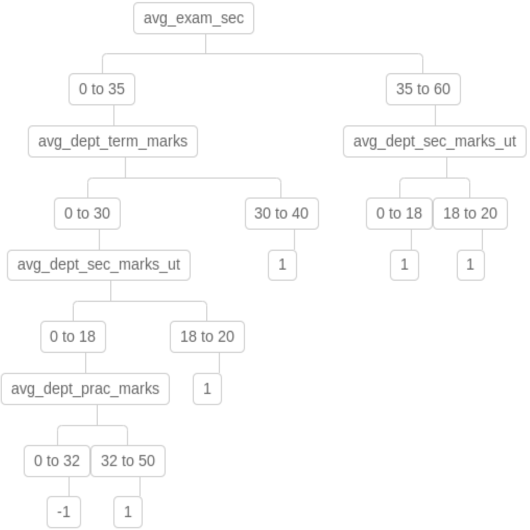


Fig.1 Decision tree for Mumbai University

Initially the raw dataset is available with attributes Student_id, Name, Theory_mrks, Practical_mrks,

DOI 10.17148/IJARCCE.2017.6216

Term_mrks, Unit_mrks in a single .csv file.Total of about 30,000 actual student academic dataset is considered. Though whole dataset is obtained in a single .csv file; it was segregated into different files for each attribute. These separate files are uploaded into the system for pre-processing which calculates the averages for each section for individual students. In the next part, algorithm is implemented. Decision tree is generated for each university and the tree rules are stored in database table named Dataset. Depending on these rules the students are also sorted and maintained in separate tables in the backend.

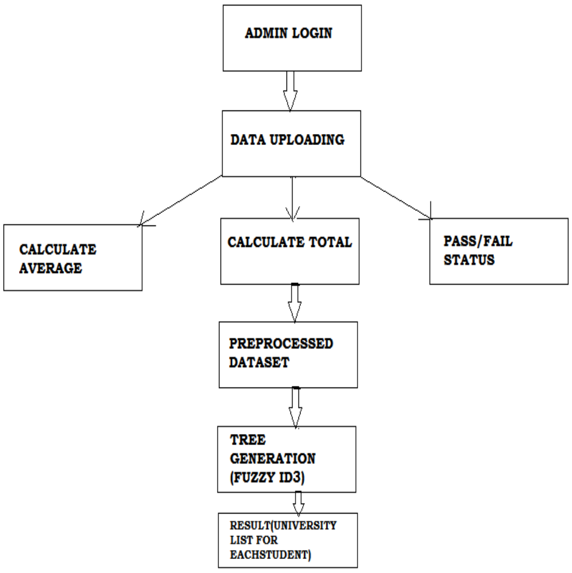
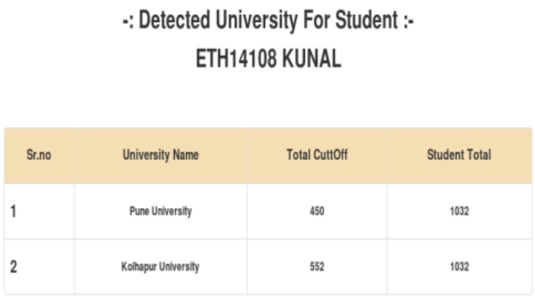


Fig.2. Flow of the System

On the GUI, on selecting the student name from the list, the name of the universities in which that particular student has the chance of getting admission is displayed. The columns displayed include name of university, University cutoff and student total. While for those students who doesn't fit in any of the universities a message "Sorry no result found" is displayed.

V. EXPERIMENTAL RESULTS

Once all the dataset has been uploaded the preprocessing is done and the resulting dataset is displayed. Fig3.shows the preprocessed dataset.



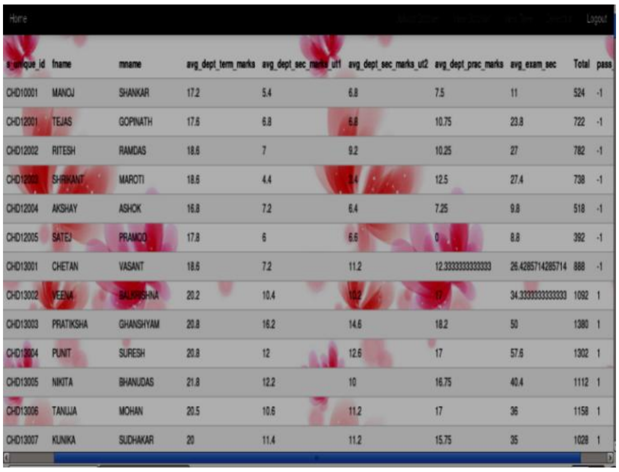
-: Detected University For Student :-

ETH14108 KUNAL

Sr.no	University Name	Total CuttOff	Student Total
1	Pune University	450	1032
2	Kolhapur University	552	1032

Fig.3. Preprocessed dataset

The Result i.e the Universities applicable for the individual students is shown in the fig.4.While for those whose score is much less; a message saying "Sorry no result found "is displayed.



s_unique_id	fname	mname	avg_dept_term_marks	avg_dept_sec_marks_ut1	avg_dept_sec_marks_ut2	avg_dept_prac_marks	avg_exam_sec	Total	pass
CHD10001	MANOJ	SHANKAR	17.2	5.4	6.8	7.5	11	524	-1
CHD12001	TEJAS	GOPINATH	17.6	6.8	6.8	10.75	23.8	722	-1
CHD12002	RITESH	RAMDAS	18.6	7	9.2	10.25	27	782	-1
CHD12003	SHRIKANT	MAROTI	18.6	4.4	3.4	12.5	27.4	738	-1
CHD12004	AKSHAY	ASHOK	16.8	7.2	6.4	7.25	9.8	518	-1
CHD12005	SATEJ	PRAMOD	17.8	6	6.6	0	8.8	392	-1
CHD13001	CHETAN	VASANT	18.6	7.2	11.2	12.3333333333333	26.4285714285714	888	-1
CHD13002	VEENA	BALKRISHNA	20.2	10.4	10.2	17	34.3333333333333	1092	1
CHD13003	PRATIKSHA	GHANSHYAM	20.8	16.2	14.6	18.2	50	1380	1
CHD13004	PUNIT	SURESH	20.8	12	12.6	17	57.6	1302	1
CHD13005	NIKITA	BHANUDAS	21.8	12.2	10	16.75	40.4	1112	1
CHD13006	TANUJA	MOHAN	20.5	10.6	11.2	17	36	1158	1
CHD13007	KUNIKA	SUDHAKAR	20	11.4	11.2	15.75	35	1028	1

Fig.4. Final output

VI. CONCLUSION

The proposed system has been designed using one of the most popular approaches for big data; MapReduce framework, which is a functional programming paradigm that is well suited to handle parallel processing of huge data sets. The decision tree technique in FID3 algorithm involves constructing a tree to model the classification process [7]. Thus, understandable prediction rules are created from the training data, which helps in classification process more easily. The output results showed that using this system the admin can get an idea about the students enrolment in next level of education. From the listed universities for a student, he/she can choose any of the university and go on with further admission process. Validation of multiple conditions and criteria improves the performance of the prediction of universities for the students according to their scores and university cut-offs. Thus, the concept of predictive analytics holds great promise for helping educational institution make evidence- based decisions related to students.

REFERENCES

[1] "Neuro-Fuzzy Classification Approach To The Assessment Of Student Performance". Arif S. Al-Hammadi and R. H. Milne Etisalat College of Engineering Emirates Telecom. Corporation Sharjah, P.O. Box: 980 United Arab Emirates. J U k 2004

[2] "Subset hood-based Fuzzy Rule Models and their Application to Student Performance Classification." Khairul A. Rasmani and Q. Shen Department of Computer Science. The University of Wales, Aberystwyth SY23 3DB, UK. 2005 IEEE

[3] "Mining Social Media Data for Understanding Students' Learning Experiences." Xin Chen, Student Member, IEEE, Mihaela Vorvoreanu, and Krishna Madhavan.

[4] "Inducing Fuzzy Models for Student Classification." Ossi Nykänen Senior Researcher, Tampere University of Technology, Digital Media Institute Hypermedia Laboratory, P.O.Box 553, FI-33101 Tampere, Finland. Educational Technology & Society.

DOI 10.17148/IJARCCE.2017.6216

[5] "Predicting Students' Performance using ID3 and C4.5 Classification algorithm". Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao Department of Computer Engineering, Fr.C.R.I.T., Navi Mumbai, Maharashtra, India International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013.

[6] "Fuzzy Rule Base System for Student grade Classification in Online Test". Solaf Hussain1 and Miran H. Mohammed Baban". International Journal of Scientific & Engineering Research, Volume 6, Issue 8, August-2015 ISSN 2229-5518

[7] "A comparative study of three Decision Tree algorithms: ID3, Fuzzy ID3 and Probabilistic Fuzzy ID3". Guoxiu Liang269167 Bachelor Thesis Informatics & Economics Erasmus University Rotterdam ,Netherlands Augustus 2005

[8] "On the use of MapReduce to build Linguistic Fuzzy Rule Based Classification Systems for Big Data."

[9] Victoria L´opez, Sara del R´ıo, Jos´e Manuel Ben´ıtez and Francisco Herrera 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)

[10] "Predictive Analytics in Higher Education." Shankar M. Pati**l** International Journal of Advanced Research in Computer and Communication Engineering

[11] Vol. 4, Issue 12, December 2015

[12] Jindal Rajni and Dutta Borah Malaya, “Predictive Analytics in Education Context” IT Pro Publ ished by the IEEE Computer Society, July/August 2015.

DOI 10.17148/IJARCCE.2017.6216