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 3, March 2017

Classification System to Predict Universities for Students using Fuzzy Logic

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: Big Data is relatively a new concept and a lot of definitions have been given to it by researchers, organizations and individuals. To extract meaningful value from big data, you need optimal processing power and analytics capabilities**.** 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, Predictive Analytics in Higher Education, 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. An educational institution needs to have an approximate prior knowledge of the students performance and admittance in future studies. Taking reference of the earlier works, the proposed system looks into the academic scores of the students and classifies the students on the basis of their scores to various universities, the system can be used to determine the chance of getting admission to particular university as well. The algorithm for building the decision tree, which is an intermediate module is ID3.The system is developed in MapReduce framework which can easily handle the processing of huge dataset.

Keywords: Classification, ID3, Big data, Fuzzy rule Hadoop, Mapreduce, Decision tree.

I. INTRODUCTION

Big data technology must support search, development, governance and analytics services for all data types—from transaction and application data to machine and sensor data to social, image and geospatial data, and more. 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[12]. 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 behavior and events."[9][10].

Big Data is the core of most predictive analytic services offered by IT organizations. By the technological advances in computer hardware i.e. faster CPUs, cheaper memory, and MPP architectures and new technologies such as Hadoop, MapReduce, and in database and text analytics for processing big data, it is now feasible to collect, analyze, and mine massive amounts of structured and unstructured data for new insights[13]. 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, data-curation, search,sharing, storage,transfer, visualization, querying an d information 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, analyse 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

Various researchers have studied related work including the development of fuzzy systems using different methods and algorithms. Few of their works which helped in implementing the proposed system are included in this section.

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.

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 3, March 2017

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]. In a study related to predictive analysis the author stated that , analytics is the process of discovering, analyzing, and interpreting meaningful patterns from large amounts of data. Following this statement and other research works, he 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]. 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]. A Neuro-Fuzzy Classification Approach to the assessment of student performance used the concept of neural network and fuzzy logic. The application was to 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]. TIEN-CHIN WANG and HSIEN-DA LEE developed a tree construction procedure to build a fuzzy decision tree from a collection of fuzzy data by integrating fuzzy set theory and entropy. It proposes a fuzzy decision tree induction method for fuzzy data of which numeric attributes can be represented by fuzzy number, interval value as well as crisp value, of which nominal attributes are represented by crisp nominal value, and of which class has confidence factor. It also presents an experiment result to show the applicability of the proposed method[3].

III. METHODOLOGY

In order 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 included in the system. This method is based on the Map Reduce framework, one of the most popular approaches for big data nowadays. By means of the Map Reduce model and its different extensions, scalability can be successfully addressed, while maintaining a good fault tolerance during the execution of the algorithms. 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. Another important aspect here is the fuzzy decision tree. 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 for the respective students. Along with this, the chance of getting admission into each of the listed university is displayed in the form of percentage. Flow of the system is as shown below.

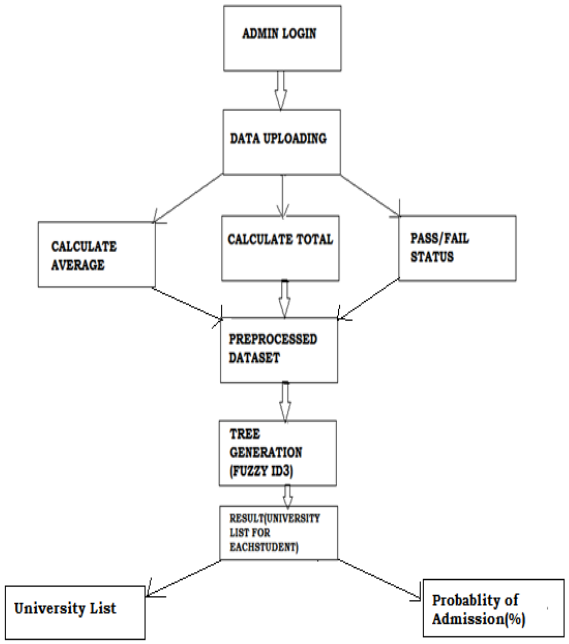


Fig.1. Flow of the proposed system

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. Many algorithms are used for building the decision tree like :

- Hunt's Algorithm (one of the earliest)
- CART
- ID3, C4.5
- SLIQ,SPRINT[11]

Of these,ID3 algorithm is the method/algorithm being implemented in my work. It builds a decision tree from some fixed or historic symbolic data in order to learn to classify them and predict the classification of new data.

ID3 uses Entropy and Information Gain to construct a decision tree. Fuzzy ID3 is only an extension of the ID3

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 3, March 2017

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 Decision Tree

Decision trees classify data by sorting them down the tree from the root to leaf nodes. Decision trees were popularized by Quinlan with the ID3 algorithm. To represent a continuous fuzzy set, we need to express it as a function and then map the elements of the set to their degree of membership. Fuzzy decision trees allow data to follow down simultaneously multiple branches of a node with different satisfaction degrees ranged on (0,1)[3]. As mentioned in the algorithm we need to follow the membership values for each attribute. We use the selected condition attribute: i.e. the attribute with highest information gain to form the decision tree. In the proposed system 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.

B. Fuzzy ID3 algorithm:

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

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

|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.

The above algorithm was implemented for tree generation. Max Gain obtained=0.18645879332542237 Overall Entropy of the dataset obtained is 0.8938591652753467

Similarly, entropy for all the attributes was calculated,

For eg. Dept_pracs

entropy=0.9448883255919571 and so on.Once the university was detected for the student, the chance of admission to that particular university was depicted by percentage.

Considering 900 marks as the total out-of mark for the students; {(cut-off total/student total) * 50}+50.

Initially the raw dataset is available with attributes Student_id, Name, Theory_mrks, Practical_mrks, Term_mrks, Unit_mrks in a single .csv file. 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 further process, algorithm is implemented.

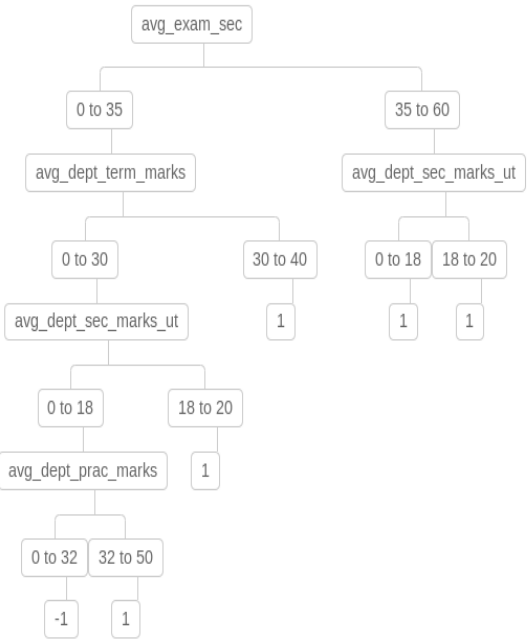


Fig.2. Decision tree for Mumbai University

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. 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 student total and probability of admission (percentage). While for those students who doesn't fit in any of the universities a message "Sorry no result found" is displayed.

V. EXPERIMENTAL RESULTS

After the dataset has been uploaded the pre-processing is done and the resulting dataset is displayed. Fig.3.shows the pre-processed dataset; with the attributes unique_id, Name, averages of Term marks, Unit test marks, Practicals, Theory marks, Total score and pass/fail status denoted by 1 and -1.

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 3, March 2017

a_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
CHD12001	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	8.4	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	13.2	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.3. Pre-processed dataset

The Result i.e. the universities applicable for the individual students is shown in the Fig.4. Also the chance of getting admission into each university is shown by the percentage; by keeping 50% as the margin. It can be read as for example student Nikita, 61.7% chance is to get admission into Pune University

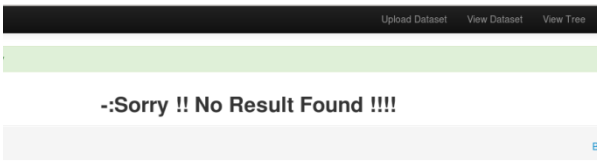
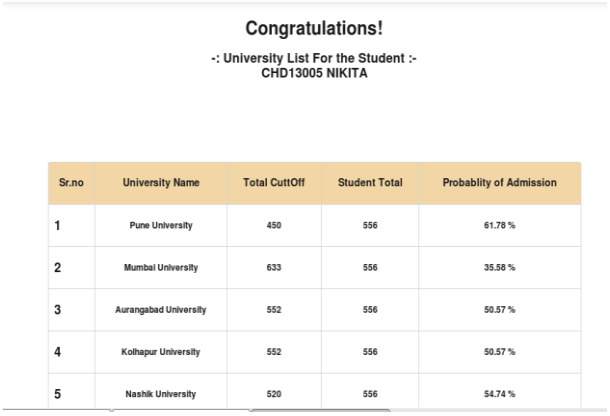


Fig. 4. Final output with Sorry message



Sr.no	University Name	Total CuttOff	Student Total	Probablity of Admission
1	Pune University	450	556	61.78 %
2	Mumbai University	633	556	35.58 %
3	Aurangabad University	552	556	50.57 %
4	Kolhapur University	552	556	50.57 %
5	Nashik University	520	556	54.74 %

Fig.5. Final output

VI. CONCLUSION

The proposed system helps the students in admission process. They can decide upon the university they wish to apply. The system has been designed using one of the most popular approaches for big data; Map Reduce 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].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. Students also gets the predicted the chance of getting admission into these universities when they make a choice. 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.

REFERENCES

[1] "Neuro-Fuzzy Classification Approach To The Assessment Of Student Performance". Arif S. Al-Hammadi and R. H. Milne Etisalat College of Engineering
[2] Emirates Telecom. Corporation Sharjah, P.O. Box: 980 United Arab Emirates. J U k 2004
[3] "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
[4] " Constructing a Fuzzy Decision Tree by Integrating Fuzzy Sets and Entropy" TIEN-CHIN WANG (王天津) HSIEN-DA LEE(李賢達)Department of Information Management-tcwang@isu.edu.tw Shou University Fortune Institute of Technology Kaohsiung, Taiwan- tcwang@isu.edu.tw
[5] "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.
[6] "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.
[7] "Fuzzy Rule Base System for Student grade Classification in Online Test". Solaf Hussain1 and Miran H. Mohammed Baban". International Journal of
[8] Scientific & Engineering Research, Volume 6, Issue 8, August-2015 ISSN 2229-5518
[9] "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
[10] "On the use of MapReduce to build Linguistic Fuzzy Rule Based Classification Systems for Big Data."
[11] Victoria L´opez, Sara del R´ıo, Jos´e Manuel Ben´ıtez and Francisco Herrera 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)
[12] "Predictive Analytics in Higher Education." Shankar M. Patil International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 12, December 2015
[13] Jindal Rajni and Dutta Borah Malaya, "Predictive Analytics in Education Context" IT Pro Published by the IEEE Computer Society, July/August 2015.
[14] "Big-data Tutorial" Marko Grobelink Jozef Stefan Institute, Slovenia. May 8th, Stavenger
[15] "Big Data: Concept, Challenges and Management Tools" Ranjana Bahri Research Scholar, Punjab Technical University, Kapurthala Assistant Professor in KCLIMT, Jalandhar, Punjab, India Vol 5, Issue 2, February 2015 ijarcsse
[16] https://www.forbes.com/sites/gilpress/2017/01/20/6-predictions-for-the-203-billion-big- data-analytics-market/#7eb0ca622083