International Journal of Computer Applications (0975 – 8887)
Volume 147 – No.5, August 2016

Effect of Numerous Data Sets on Performance Prediction

Jyoti Upadhyay
Research Scholar
AISECT University
Bhopal, M.P.

Pratima Gautam, PhD
Dean IT Dept.
AISECT University
Bhopal, M.P.

ABSTRACT
There are many factors, which may affects performance. But the number of factors also affects on result of computational model. We are presenting a computational model to forecast students' performance. To calculate we will use 8 different factors that are directly or indirectly influence performance. Influencing factors how much correlated to each other we also present this. Through this paper we classify those factors using fuzzy decision tree.

Keywords
Data Mining, Fuzzy Decision Tree, Classification, correlation and Prediction.

1. INTRODUCTION
1.1 Educational Data Mining
Data mining is a knowledge mining technique. We may refer it with knowledge discovery process. Data mining provides different techniques to mine information from the database. If user wants to mine information from educational data from then it becomes Educational Data Mining (EDM).

According to www.educationalmining.com_"Educational

Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn".

We have lots of data mining algorithms and techniques to analyze data. Some techniques are Classification, Clustering, Regression, Artificial intelligence, Neural networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. For our work we will use classification technique.

1.2 Classification
Most commonly data-mining techniques are Classification. Classification is used to develop a model that can classify the data set. A classification technique commonly works through decision tree or neural network-based classification algorithms. Data classification process involves mainly two steps learning and classification. In Learning step the training set are analyzed by classification algorithm [1]. After applying model test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For our work we are going to work upon students' data to predict success possibility in college exam and factors, which may affect students' performance. Some classification models are:

- Decision tree induction
- Bayesian Classification
- Neural Networks
- Classification Based on Association

1.3 Decision tree
A decision tree is a classification technique, which classifies data set into possible classes. Decision trees are used to extract knowledge by making decision rules from the large amount of available information [2]. In decision tree each leaf node represents a decision and branch node represents a choice of alternatives attributes. From root node, users split each node recursively according to decision tree learning algorithm. The final result of decision tree is a possible scenario of decision and its result.

In our paper we will take fuzzy decision tree. Fuzzy decision trees grow in a top-down way. It uses recursively partitioning the training data into segments with similar outputs.

To construct Fuzzy decision tree apply fuzzy sets to describe (quantify) attributes and then use the ID3 approach. Fuzzy entropy, information gains or gain ratio are used as a measure of attribute selection [8].

2. METHODOLOGY
2.1 Data Collection
For data collection we select an educational organization. We collect 130 data according to our need. For experiment we will take following parameter.

Table 1: Parameters

S. No	Parameter	Values
1	Cast Category	SC, ST, OBC, GEN
2	Previous Percent	Percentage of 12^{th}
3	Attendance percent	Attendance Percentage
4	Location	Rural,SemiUrban, Urban
5	Sports interest	Avg, Good, Poor
6	Parents income	Avg, Good, Poor
7	Unittest Performance	Avg, Good, Poor
8	Recent Result	Fail, Pass

Above Parameter list Recent Result is label and rest are decision parameter.

2.2 Process

We filter missing data using excel feature. We have to transform data set into fuzzy data set. Fuzzy set theory is an approach to represent uncertainty. A fuzzy set A is characterized by its membership function (MF). MF range is belonging to the unit interval. MF represents the degree of membership of the point in the set A [3]. We transformed data set into fuzzy set. Transformed data set shows in table 2.

Table 2Transformed Data

Nmae of students	Cast	Location	recent result	Fuzzy attn	fuzzy prev result	sports Interest	parents_income	Unit_test Perf
Sumeet Singh	GEN	Urban	Fail	good	thierd	avg	AVG.	Poor
Sweta Prasad	ST	Semi-urban	Fail	good	second	good	good	Avg
Neelam Singh	GEN	Urban	Pass	good	second	poor	good	Good
B.Madhur Kumar	GEN	Rural		AVG	thierd	poor	poor	poor
Karan Arora	GEN	Urban	Pass	good	second	good	good	Good
Akanksha Singh	GEN	Urban	Fail	AVG	second	poor	good	poor
Gaurav Gayali	GEN	Semi-urban	Pass	good	second	good	AVG.	good
Shainy Joseph	GEN	Urban		good	first	good	poor	Good
Kamal Narayan	SC	Rural	Fail	good	second	good	AVG.	Poor

After Transformation we apply filtration on data set. Then calculate weight of each attribute. We depicted it on Figure 1. It shows that Unit test performance and attendance has high weight comparing to other. Cast has lowest weight. We use Chi square test to calculating weight.

attribute	weight
Cast	1.725
Sports Interest	5.515
fuzzy prev result	13.261
parents_income	13.692
Location	18.379
Unit test Perf	42.882
Fuzzy attn	45.302

Figure 1: weight of attribute

After Calculating weight apply validation and CHAID to construct Decision tree. The condensation of CHAID is Chi-squared Automatic Interaction Detector. It is tree classification methods originally proposed by Kass (1980). It has been developed for categorical variables[5].

CHAID will "build" non-binary trees (i.e., trees where more than two branches can attach to a single root or node), based on a relatively simple algorithm that is particularly well suited for the analysis of larger datasets. CHAID algorithm will often effectively yield many multi-way frequency tables (e.g., when classifying a categorical response variable with many categories, based on categorical predictors with many classes).

CHAID is a technique that recursively splits a data set into discrete segments. These are nodes. Nodes are split in such a way that the difference of the resultant variable is minimized [6].

To construct Decision tree we have to select root node from the list of decision attributes. Attribute selection has been performed by information gain. Basically information gain can be defined by

Information gain = Information before splitting – Information after splitting.

Information gain is influenced about choosing attributes with a large number of values. This may result in over fitting. So that modifications in information gain is Gain ratio. Gain ratio takes number and size of branches into account when choosing a decision attribute. It corrects the information gain by taking the intrinsic information of a split into account [4].

2.3 Tools

For our work we selected Rapid miner tool. Rapid Miner is an open source to maximize the value of predictive analytics. Through Rapid Miner we can makes predictive analytics lightning-fast, drastically reducing the time to discover risks. It provides powerful visual design evaluation speeds repetitive tasks, considerably reducing the time that spent on data access and preparation. Rapid miner helps connect to any data at any scale and desirable Optimization schemes [7].

3. OUTCOMES

Generated decision tree shows in Figure 2. We can See from Figure 2 and Figure 3 that if Attendance of a candidate Average then most probable performance will be poor specially when candidate from rural or semi urban area. If Attendance is good then performance value will be increases. Figure 3 shows fuzzy performance rule, extract from decision tree. All parameters are not plays vital role in performance prediction it can be easily determine by decision tree.

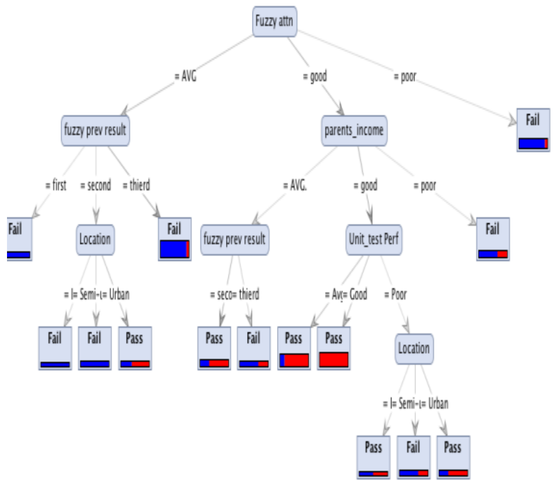

Figure 2 fuzzy decision trees

Tree

```
Fuzzy attn = AVG
|   fuzzy prev result = first: Fail {Fail=4, Pass=0}
|   fuzzy prev result = second
|   |   Location = Rural: Fail {Fail=2, Pass=0}
|   |   Location = Semi-urban: Fail {Fail=3, Pass=0}
|   |   Location = Urban: Pass {Fail=2, Pass=3}
|   fuzzy prev result = thierd: Fail {Fail=24, Pass=2}
Fuzzy attn = good
|   parents_income = AVG.
|   |   fuzzy prev result = second: Pass {Fail=2, Pass=4}
|   |   fuzzy prev result = thierd: Fail {Fail=2, Pass=1}
|   parents_income = good
|   |   Unit_test Perf = Avg: Pass {Fail=3, Pass=14}
|   |   Unit_test Perf = Good: Pass {Fail=0, Pass=20}
|   |   Unit_test Perf = Poor
|   |   |   Location = Rural: Pass {Fail=1, Pass=1}
|   |   |   Location = Semi-urban: Fail {Fail=2, Pass=1}
|   |   |   Location = Urban: Pass {Fail=1, Pass=2}
|   parents_income = poor: Fail {Fail=4, Pass=2}
Fuzzy attn = poor: Fail {Fail=11, Pass=1}
```

Figure 3: Decision rule

In Figure 4 shows predicted result and confidence value of performance. We categorized performance in fail and pass so that we can see in first row that original and predicted result is same so that confidence of fail is greater than confidence of pass.

ExampleSet (112 examples, 4 special attributes, 8 regular attributes) Filter (112 / 112 examples): all

Row No.	recent result	prediction(rece...	confidence(Fail)	confidence(Pass)	Nmae of st...	Cast	Location	Fuzzy attn	fuzzy pre
1	Fail	Fail	0.667	0.333	Sumeet Sing	GEN	Urban	good	thierd
2	Fail	Pass	0.176	0.824	Sweta Prasa	ST	Semi-urban	good	second
3	Pass	Pass	0	1	Neelam Sing	GEN	Urban	good	second
4	Pass	Pass	0	1	Karan Arora	GEN	Urban	good	second
5	Fail	Pass	0.400	0.600	Akanksha Si	GEN	Urban	AVG	second
6	Pass	Pass	0.333	0.667	Gaurav Gay	GEN	Semi-urban	good	second
7	Fail	Pass	0.333	0.667	Kamal Nara	SC	Rural	good	second
8	Pass	Pass	0.333	0.667	Arvind Tanc	SC	Urban	good	thierd
9	Pass	Pass	0.333	0.667	Shweta Sing	GEN	Urban	good	first
10	Pass	Pass	0	1	Deepali Pat	GEN	Urban	good	second
11	Fail	Fail	0.923	0.077	Vijay Phulw	GEN	Semi-urban	AVG	thierd
12	Fail	Fail	0.917	0.083	G.Megha	GEN	Rural	poor	thierd
13	Fail	Fail	0.917	0.083	Ramneek B	SC	Rural	poor	thierd
14	Pass	Pass	0.176	0.824	D.Sneha	SC	Urban	good	second
15	Fail	Fail	0.923	0.077	G.Mahesh K	GEN	Urban	AVG	thierd
16	Pass	Pass	0.500	0.500	Shubham Sc	SC	Rural	good	thierd
17	Fail	Fail	1	0	Prashant Ku	SC	Semi-urban	AVG	second
18	Fail	Fail	0.917	0.083	Dhriti Pande	GEN	Urban	poor	thierd
19	Fail	Fail	0.917	0.083	Sanju Tandi	SC	Rural	poor	thierd
20	Fail	Fail	1	0	Jatinder Sin	GEN	Semi-urban	AVG	second

Figure 4 Predicted result and Confidence Value

Figure 5 shows Performance Vector, which shows accuracy of model is 99%and classification error is 0.89%.

PerformanceVector

```
PerformanceVector:
accuracy: 99.11%
ConfusionMatrix:
True:   Fail    Pass
Fail:   61      1
Pass:   0       50
classification_error: 0.89%
ConfusionMatrix:
True:   Fail    Pass
Fail:   61      1
Pass:   0       50
kappa: 0.982
ConfusionMatrix:
True:   Fail    Pass
Fail:   61      1
Pass:   0       50
weighted_mean_recall: 99.02%, weights: 1, 1
ConfusionMatrix:
True:   Fail    Pass
Fail:   61      1
Pass:   0       50
weighted_mean_precision: 99.19%, weights: 1, 1
ConfusionMatrix:
True:   Fail    Pass
Fail:   61      1
Pass:   0       50
```

Figure 5: Performance Vector

4. CONCLUSION

And Fuzzy Decision rule will be extract like in Figure 3 Decision tree shows that attendance is play an important role in students' performance. Students' performance may affect by living location and parents income also. Cast has no role in performance. Organization, Teachers and students also can take advantage from this model. They can find out factor which may affect result and do work on it. Different factors may help to predict performance. If through a factor it is difficult to predict performance then other factor my help for prediction. We perform this task for 150 samples. After filtration we got 112 samples in which only 1 produce wrong prediction.

5. REFERENCES

[1] Ramageri, Mrs. Bharati M. "DATA MINING TECHNIQUES AND APPLICATIONS ." Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305 (2010): 301-305.

[2] Sonia Joseph, Laya Devadas. "Student's Performance Prediction Using Weighted Modified ID3 Algorithm ." International Journal of Scientific Research Engineering & Technology (IJSRET), 4.5 (2015): 571.

[3] Zadeh, L. A. "fuzzy set." information and control 8 (1965): 338-352.

[4] Alec Pawling, Nitesh V. Chawla, and Amitabh Chaudhary. "Computing Information Gain in Data Streams." TDM 2005: 2005 Temporal Data Mining Workshop (2005): 72-81.

[5] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data", Applied Statistic, Vol. 29, (1980): 119-127.

[6] M. Ramaswami , R. Bhaskaran , "A CHAID Based Performance Prediction Model in Educational Data Mining",International Journal of Computer Science Issues, Vol. 7, (2010) :10-18

[7] https://rapidminer.com/products/studio/, dated march 2016.

[8] Xiaodong Liu,Xinghua Fenga, andWitold Pedryczc, "Extraction of fuzzy rules from fuzzy decision trees: An axiomatic fuzzy sets (AFS) approach", Data & Knowledge EngineeringVolume 84, (2013) , 1–25