Application of Decision Tree in Analysis of Intra College Festival Data set

[1]Sumit Srivastava (Dr.)
[1]Department of Information Technology,
Poornima College of Engineering, Jaipur, Rajasthan-302022(India)

[2]Megha Yadav
[2]Department of Computer Science,
Poornima College of Engineering, Jaipur, Rajasthan-302022(India)

[3]Sandeep Devra
[3]Department of Computer Science,
Jagannath Institute of Engineering & Technology, Jaipur, Rajasthan-302022(India)

Abstract - Data mining has been used very frequently to extract hidden information from large databases. The paper suggested the use of data mining technique named decision trees along with its algorithm for continuously monitoring the College fest. In this way the data extracted could be used to teach the students on the real time problem scenario application to be monitored at college level. Also the model can be used for the future planning of such fest at college level. The algorithm used is CHAID (Chi-squared Automatic Interaction Detection) using PASW18 as software.

Keywords- Data Mining, ID3, C4.5, CART, CHAID

1. Introduction

It is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. It is an increasingly important tool by modern business to transform data into business intelligence. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. In this paper, the effective use of the information available by using the decision trees on a festival data set. The problem is to classify the information available on the basis of year of student participated in the various events, participation on the basis of the gender, particiaption on the basis of type of event(can be clutural, club, sports and tehnical), branch wise information. The algorithm available to be used in Decision trees ID3, C4.5, CART, CHAID to be used with software tools like PASW18 onwards. Also the comparison with the other algorithms is being done.

2. Decision Trees

Decision trees are often used in classification and prediction. It is simple yet a powerful way of knowledge Representation [1]. The models produced by decision trees are represented in the form of tree structure. A leaf node indicates the class of the examples. The instances are classified by sorting them down the tree from the root node to some leaf node.

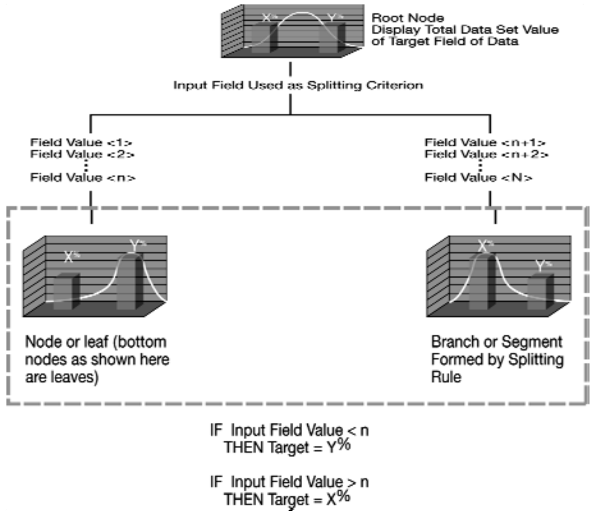

Root Node
Display Total Data Set Value
of Target Field of Data
Input Field Used as Splitting Criterion
Field Value <1>
Field Value <2>
Field Value <n>
Field Value <n+1>
Field Value <n+2>
Field Value <N>
X
Y
Node or leaf (bottom nodes as shown here are leaves)
Branch or Segment Formed by Splitting Rule
IF Input Field Value < n THEN Target = Y%
IF Input Field Value > n THEN Target = X%

FIG 2.1: Decision tree reflecting both continuous and categorical object of analysis.

The reason of using Decision trees are

1. Decision trees are white boxes means they generate simple, understandable rules
2. Decision trees are non-parametric means no specific data distribution is necessary.

3. Decision trees handle missing values as easily as any normal value of the variable.
4. In decision trees elegant tweaking is possible.
5. Decision trees identify subgroups. Each terminal or intermediate leave in a decision tree can be seen as a subgroup/segment of your population.
6. Decision trees run fast even with lots of observations and variables
7. Decision trees can be used for supervised and unsupervised learning.
8. Decision trees can easily handle unbalanced datasets.
9. Versatility for a wide variety of data mining tasks, such as classification, regression, clustering and feature selection

while some of the disadvantage lies as following

1. While decision trees classify quickly, the time for building a tree may be higher than another type of classifier.
2. Decision trees suffer from a problem of error propagation
3. Disjunctive descriptions may be required
   a. naturally represent disjunctive expressions

the algorithms being used while decision tree analysis is as follows.

2.1 ID3 Algorithm:- it is a greedy search technique, for which top-down recursive divide and conquer manner. Generally the question relates to classifying the information. It generally use information gain as a quantitative measure of an attribute[3]. The algorithm for attribute selection measure

- Select the attribute with the highest information gain
- Let p_i be the probability that an arbitrary tuple in D belongs to class C_i, estimated by $|C_{i,\,D}|/|D|$
- Expected information (entropy) needed to classify a tuple in D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad \ldots\ldots(2.1)$$

- Information needed (after using A to split D into v partitions) to classify D:
    - o Information gained by branching on attribute A (If A is discrete valued):

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j) \qquad \ldots\ldots..(2.2)$$

- Information gained by branching on attribute A (If A is discrete valued):

$$Gain(A) = Info(D) - Info_A(D) \qquad \ldots\ldots\ldots(2.3)$$

2.2 C4.5 Algorithm:- C4.5 (a successor ofID3) uses gain ratio to overcome the problem. Here for Computing the information- gain for continuous value attributes the best spilt-point is being calculated [2][4]. The algorithm for calculating the same is:

- Let attribute A be a continuous-valued attribute
- Must determine the *best split point* for A
- Sort the value A in increasing order
- Typically, the midpoint between each pair of adjacent values is considered as a possible *split point*
- $(a_i+a_{i+1})/2$ is the midpoint between the values of a_i and a_{i+1}
- The point with the *minimum expected information requirement* for A is selected as the split-point for A

and the SPLIT is calculated as

- D1 is the set of tuples in D satisfying A ≤ split-point, and D2 is the set of tuples in D satisfying A > split-point

It uses the kind of normalization to information gain using a "Split Information" value defined analogously with Info(D) as:

$$\boxed{SplitINFO_A(D) = -\sum_{i=1}^{k} \frac{D_i}{D} \log \frac{D_i}{D}} \qquad \ldots\ldots..(2.4)$$

and the gain ratio defined as

$$\boxed{GainRATIO(A) = \frac{GAIN(A)}{SplitINFO(A)}} \qquad \ldots\ldots\ldots\ldots(2.5)$$

2.3 CART Algorithm: Classification and Regression Trees (CART) is a non-parametric technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. The Gini index is used in CART. Gini index measures the impurity of *D, a data partition or set of training tuples, as*

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2 \qquad \ldots(2.6)$$

where p_j is the relative frequency of class j in D. The Gini Index consider the binary split of the each of the attribute. If the data set D is split into subset D_1 and D_2, the gini-index can be defined as

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2) \quad \text{...(2.7)}$$

and Reduction in Impurity (that would be incurred by a binary split on a discrete- or continuous-valued attribute A) is:

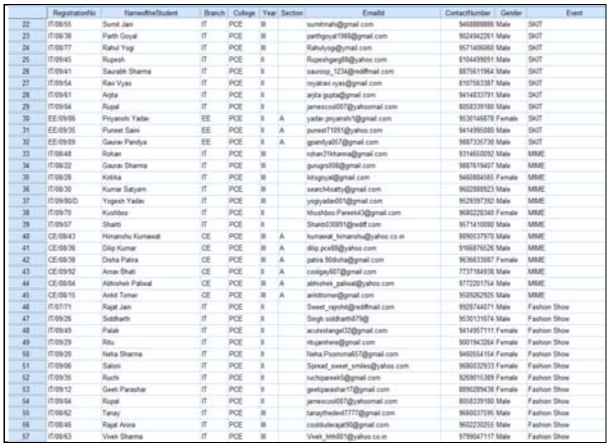
$$\Delta gini(A) = gini(D) - gini_A(D) \quad \text{....(2.8)}$$

2.5 CHAID Algorithm:- Chi-squared Automatic Interaction Detection is a non-binary decision tree algorithm, where split search is designed for categorical values. CHAID uses a Chi-squared test using contingency tables. The advantage for using CHAID as method of evaluation is:

- Used in fields of marketing research and public health
- Can handle a large number of variables
- Designed to identify potentially meaningful patterns in a dataset
- Easy to use and understand

3. Decision Trees

The application data set consists of the cultural fest named "AAROHAN2k10" organized as annual fest in one of the private engineering colleges. It includes nearly 22 technical events, 13 club events, 7-8 sports events, 9-10 cultural events with 5500 approx student participating for the competition. For each student participation in activity involving Technical and club is a must with total of 4 events per student has to be submitted. So with almost 5500*4= 22,000 entries as input the event has been successfully completed. The snapshot of the data taken from AAROHAN2k10 is shown below



	RegistrationNo	NameoftheStudent	Branch	College	Year	Section	EmailId	ContactNumber	Gender	Event
22	IT/08/55	Sumit Jain	IT	PCE	III		sumitmahi@gmail.com	9468888886	Male	SKIT
23	IT/08/38	Parth Goyal	IT	PCE	III		parthgoyal1988@gmail.com	9024942261	Male	SKIT
24	IT/08/77	Rahul Yogi	IT	PCE	III		Rahulyogi@ymail.com	9571406060	Male	SKIT
25	IT/09/45	Rupesh	IT	PCE	II		Rupeshgarg88@yahoo.com	8104499091	Male	SKIT
26	IT/09/41	Saurabh Sharma	IT	PCE	II		saurosp_1234@rediffmail.com	8875611964	Male	SKIT
27	IT/09/54	Ravi Vyas	IT	PCE	II		royalravi.vyas@gmail.com	8107563387	Male	SKIT
28	IT/09/61	Arpita	IT	PCE	II		arpita.gupta@gmail.com	9414833791	Male	SKIT
29	IT/09/04	Rupal	IT	PCE	II		jamescool007@yahoomail.com	8058339180	Male	SKIT
30	EE/09/06	Priyanshi Yadav	EE	PCE	II	A	yadav.priyanshi1@gmail.com	9530146878	Female	SKIT
31	EE/09/35	Puneet Saini	EE	PCE	II	A	puneet71091@yahoo.com	9414995080	Male	SKIT
32	EE/09/09	Gaurav Pandya	EE	PCE	II	A	gpandya057@gmail.com	9887335730	Male	SKIT
33	IT/08/48	Rohan	IT	PCE	III		rohan31khanna@gmail.com	9314650092	Male	MIME
34	IT/08/22	Gaurav Sharma	IT	PCE	III		gurugrs008@gmail.com	9887619407	Male	MIME
35	IT/08/28	Kritika	IT	PCE	III		kitsgoyal@gmail.com	9460884565	Female	MIME
36	IT/08/30	Kumar Satyam	IT	PCE	III		search4satty@gmail.com	9602888923	Male	MIME
37	IT/09/80/D	Yogesh Yadav	IT	PCE	III		yogiyadav001@gmail.com	9529397392	Male	MIME
38	IT/09/70	Kushboo	IT	PCE	II		khushboo.Pareek43@gmail.com	9680228340	Female	MIME
39	IT/09/07	Shakti	IT	PCE	II		Shakti030891@rediff.com	9571410080	Male	MIME
40	CE/08/43	Himanshu Kumawat	CE	PCE	III	A	kumawat_himanshu@yahoo.co.in	8890537970	Male	MIME
41	CE/08/36	Dilip Kumar	CE	PCE	III	A	dilip.pce88@yahoo.com	9166876526	Male	MIME
42	CE/08/38	Disha Patra	CE	PCE	III	A	patra.90disha@gmail.com	9636633087	Female	MIME
43	CE/09/92	Aman Bhati	CE	PCE	II	A	coolgay607@gmail.com	7737184938	Male	MIME
44	CE/08/04	Abhishek Paliwal	CE	PCE	III	A	abhishek_paliwal@yahoo.com	9772201754	Male	MIME
45	CE/08/15	Ankit Tomar	CE	PCE	III	A	ankittomer@gmail.com	9509262925	Male	MIME
46	IT/07/71	Rajat Jain	IT	PCE	II		Sweet_rajrohit@rediffmail.com	9928744071	Male	Fashion Show
47	IT/09/26	Siddharth	IT	PCE	II		Singh.siddharth879@	9530131074	Male	Fashion Show
48	IT/09/49	Palak	IT	PCE	II		acutestangel32@gmail.com	9414957111	Female	Fashion Show
49	IT/09/29	Ritu	IT	PCE	II		ritujainhere@gmail.com	9001943264	Female	Fashion Show
50	IT/09/20	Neha Sharma	IT	PCE	II		Neha.Poornima657@gmail.com	9460554154	Female	Fashion Show
51	IT/09/06	Saloni	IT	PCE	II		Spread_sweet_smiles@yahoo.com	9680032933	Female	Fashion Show
52	IT/09/35	Ruchi	IT	PCE	II		ruchipareek5@gmail.com	9269015389	Female	Fashion Show
53	IT/09/12	Geeti Parashar	IT	PCE	II		geetiparashar17@gmail.com	8890289438	Female	Fashion Show
54	IT/09/04	Rupal	IT	PCE	II		jamescool007@yahoomail.com	8058339180	Male	Fashion Show
55	IT/08/62	Tanay	IT	PCE	III		tanaythedevil7777@gmail.com	9660037595	Male	Fashion Show
56	IT/08/46	Rajat Arora	IT	PCE	III		cooldudearajat90@gmail.com	9602230255	Male	Fashion Show
57	IT/08/63	Vivek Sharma	IT	PCE	III		Vivek_hhh001@yahoo.co.in	9799047117	Male	Fashion Show

Fig 3.1: Training Data Set (taken from Aahoran database)

During the result analysis, it was asked to predict the participation in each types of events by each branch of each college with Gender of the students can be as the dependent variable for the analysis. Also if the event is supposed to a group then classification of the information on the basis of college, year, branch, team-name, event-name is requested. For such problem CHAID is selected as the best technique for the solution as information can be branched using non-binary variable. Some of the results snapshots calculated using PASW 18 beta version as software is shown below.

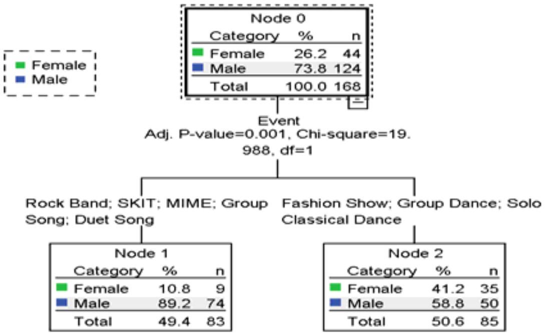


Fig 3.2:- When the dependant variable is taken to be Gender and independent variables are: Event, College,

Branch)

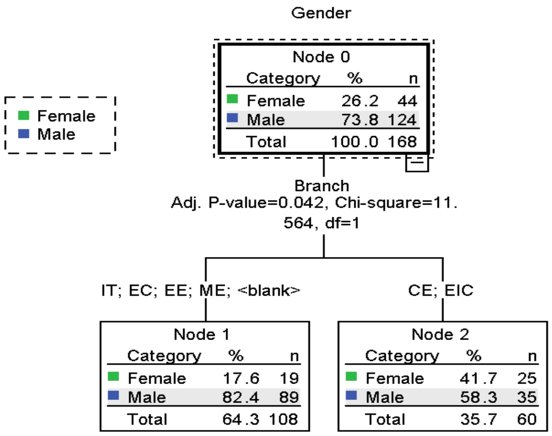


Fig 3.3:- When the dependant variable is taken to be Gender and independent variables are: College, Branch)

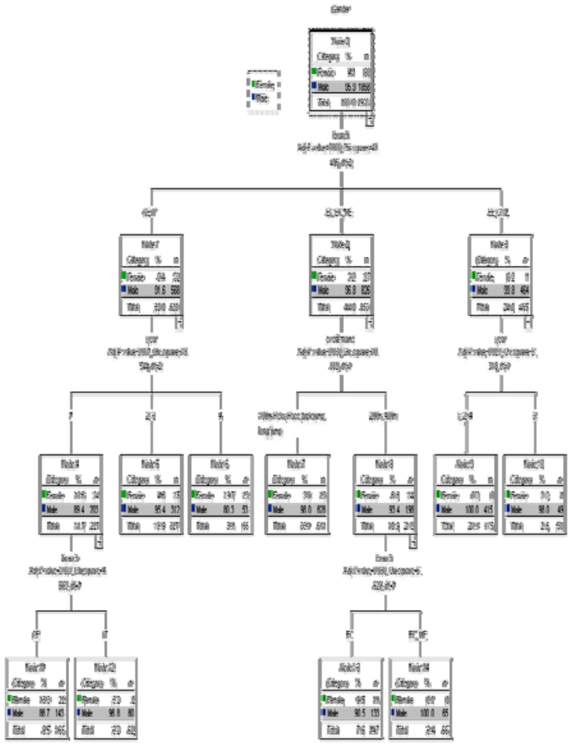


Fig 3.4:- When the dependant variable is taken to be Gender and independent variables are: college, year, branch, team_name, event_name

4. Conclusion

The inference from the various result analyses can be shown as.

- It was predicted that 89.2% out of 73.9% of the males have participated in events like Rock Band, SKIT, MIME, Group Song and Duet Song.
- On the basis of the data studied it can be concluded that 58.8% out of 73.9% of the males have participated in events Cultural and Club Events
- it can be further concluded that 58.3% out of 73.8% of the males are from CE, EIC branches who have participated in cultural events
- On the basis of the data studied it can be concluded that 82.4% out of 73.8% of the males are from IT, ME, EC, EE branches who have participated in cultural events, sports and technical events

With fast analysis of large databases and classification of information using Decision trees, CHAID is founded as the one among the best algorithm for the large data analysis.

5. Acknowledgement

We acknowledgement the Director of the engineering institute for providing us the opportunity to work on their database.

6. Bibliographies and References

[1] T. Mitchell, *"Decision Tree Learning*", in T. Mitchell, Machine
Learning, the McGraw-Hill Companies, Inc., 1997, pp. 52-78.
[2] P. Winston, "*Learning by Building Identification Trees*", in P. Winston, Artificial Intelligence, Addison-Wesley Publishing Company, 1992, pp. 423-442.
[3] Howard J. Hamilton's CS Course: Knowledge Discovery in Databases. Accessed 06/06/12.
[4] http://www.cs.waikato.ac.nz/ml/weka/, accessed 06/05/21.
[5]http://grb.mnsu.edu/grbts/doc/manual/J48_Decision_Trees.html, accessed 06/06/12.
[6] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kauffman, 1993.

[1]Dr. Sumit Srivastava
Professor,
Department of Information Technology,
Poornima College of Engineering,
Jaipur-302022(raj)

[2]Mrs. Megha Yadav
M.Tech, Research Scholar,
Department of Computer Science Engineering,
Poornima College of Engineering,
Jaipur-302022(raj)

[3]Mr.Sandeep Devra
Associate Professor,
Department of Computer Engineering,
JNIT, Jaipur-302022(Raj)