Chapter 5: A Case Study in Peer Learning using ID3

In this chapter, a methodology based on ID3 (a decision tree based learning technique) for classification of disabled students of IGNOU has been evolved, for creating peer-to-peer student groups. This can facilitate an environment for mutual help and academic interaction. The purpose is to promote blended social learning. Section after introduction provide information about the dataset and variables followed by classification, decision trees, ID3, data analysis, results and discussion about these results, respectively. This case study can be called as start of a Peer Recommender System (PRS). As per sections 2.16 & 4.3, this study starts from scratch (‘0’) i.e. a cold start.

5.1 Introduction

Data Mining or Pattern Finding is one of the established areas of Computer Science. Data Mining is a confluence of various other disciplines: artificial intelligence, statistics, databases, machine learning, visualization etc. Further it involves knowledge of the domain of application viz. finance, medicine, industries, education. Educational data mining is a new domain where data obtained from educational institutions is used to analyse and find interesting patterns. This study is based on grouping of disabled students of IGNOU for the purpose of more effective peer-to-peer learning using ID3[49].

Initiating grouping of students for creating better learning environments is a problem. It is an even bigger problem when it is concerned with students with special need, in the institution. Such deprived learners include – disabled students, females, economically poor, backward classes, socially deprived, employed and those living in remote areas etc. The number of disabled students in a class is generally very small and hence a group of similar students is not available with whom they can comfortably interact. If a group can be formed then, these students can of course interact amongst themselves and discuss many issues they face which are not faced by normal students. Forming of groups can be more useful in a
distance learning scenario when students meet only for two days in a week i.e. Saturdays and Sundays. The normal students here are mostly working or have personal responsibilities or may be both. These points can limit the possibility of interactions. A solution to this is to bring these students in contact with a similar peer. In this paper, there is a grouping method described using real world data.

5.2 About SOCIS Data Sets

For classification, data or profile of disabled students is obtained from SRD of IGNOU for two cycles or semesters of the year 2009. This data was obtained by IGNOU from students filling of the enrolment forms. Out of this data students who joined School of Computer and Information Sciences (SOCIS) in both cycles are selected. SOCIS offers four distance learning courses.

- MCA (Master of Computer Applications)
- BCA (Bachelor of Computer Applications)
- CIT (Certificate in Information Technology)
- CIC (Certificate in Computing)

Figure 46: Most disabled students are enrolled in BCA course of SOCIS, IGNOU
The dataset is highly detailed and there are many errors. After data cleaning, following attributes are selected for grouping students:

1. Area
2. Handicap Nature
3. Region

Students in with same/similar attributes can be put in one group. This facilitates group learning. This is in accordance with blended learning where students in same group communicate online as well as offline. This justifies the choice of attributes as students with same handicap nature have similar problems and solutions of those problems. Students from same area and type of region may again encounter similar hurdles.

5.3 Variables from datasets

There is only a small fraction of total students who are disabled and who enter the university system, and the data can be grouped according to three attributes each having some attributes values – Area (Urban, Rural, Tribal) i.e. A1, B2, C3; Region (Central, Delhi, East, West, North, South) i.e. C, D, E, W, N, S; Handicap Nature or A1, B2, C3, D4, E5; Type of Handicap (Locomotor Impairment, Low Vision, Visual Impairment, Speech and Hearing Impairment, Any other Please specify) respectively. These attributes have been selected as per the mandate of the university.

General learners settle in groups easily and can manage through activities required for the course, but disabled learners, who are just one or two in a study centre, can’t be easily grouped for interaction and other activities. This study makes creation of groups all the more useful in the case of disabled students. The peer learning many facilitate motivation & joy for effective learning.
5.4 Classification

Classification techniques can be used to predict academic success, course outcomes. Approaches used for classification are: decision trees. Bayesian classifiers, neural networks, nearest neighbour classifiers, support vector machines, and linear regression. Based on CLS (section 2.5), ID3 (Iterative Dichotomiser 3) is a Decision Tree based classification technique, invented by J. Ross Quinlan in 1986[49]. It is based on the concept of entropy. Entropy is a measure of disorderness or randomness in data. Here, it is a measure of uncertainty or indetermination for a piece of information.

5.5 Decision Trees

![Decision Tree Diagram](image)

Decision trees are a data structure oriented data mining tool. They look like a tree shaped graph. As the name suggests, they enable decision making in various fields and studies like engineering, industry, medical research, education and psychology etc. A Decision Tree gives rules on the basis of which decisions are made and added to a knowledge base. Figure 47 is an example of a DT. The first node is called root node (age in this case). First or root node is the best classifier. This is level 0. This yields in 4 branches which are
pre-decided categories namely Age>55, 36<Age<55, 18<Age<35, Age<18. Decisions are possible when Age>55 and when 18<Age<35 and tree growth stops at these two branches. But at the other two branches more attributes are needed to decide. For Age<18, Income reveals maximum information and for 36<Age<55 Marital Status yields maximum information. DTs based on real world data are generally bigger.

5.6 ID3

ID3 is a decision tree [72] based data mining technique where an expert system learns from rules obtained from the data. This technique is a like tuple analysis done on attributes and is also similar to sorting, when done manually on small datasets (in register). In the process, groups or sets may also be formed. This technique decides the importance of attributes on the basis of Entropy. Entropy is a measure of disorderness or randomness in data. Here, it is a measure of uncertainty or indetermination for a piece of information ‘a’ as:

number of bits, \( n = - \log_2 P(a) \)

**Derivation of the formula**

A variable X has k possible values, all equally probable. As its n instances i.e. \( X_1, X_2 \ldots X_n \) occur, the probability of a combination of all \( X_i \)'s is defined as \( P(X_1X_2 \ldots X_n) = 1/(k^n) \)

i.e. \( k^n = 1/P(X) \)

or \( n \log(k) = - \log(P(X)) \) (taking logarithm with base k throughout, using \( \log_k k = 1 \))

\( n = - \log(P(X)) \)

where n is a measure of uncertainty or indetermination for a piece of information i.e. combination or set \( X_1X_2 \ldots X_n \). It is called Entropy. The unit of Entropy is ‘bit’ [101].
Consider weighted mean of all n, entropy E of a random variable X(w) with a probability distribution P(X(w)=x) is defined by:

\[ E = - \sum P(X(w)=x) \cdot \log_2(P(X(W)=x)) \text{ for all } x \] [21].

Conventionally, \( 0 \cdot \log_2 0 = 0 \) (using concept of limits)

### 5.7 Data Analysis and Results

This concept can be used to partition huge data into clusters or sets. Here 59 records, without wrong values and blanks are selected. If entropy decreases after partitioning then information is gained. The target here is, to find, which attribute can give the maximum gain. First step is to calculate the entropy of the present system, i.e. E(0) at level-0. Next, gain is found for partitioning by Region, Area and Handicap Nature one by one. This is level-1.

For partitioning by Region, number of students who enrolled in MCA, BCA, CIT, and CIC is counted for every value of the attribute area. Entropy for partitioning by a variable is weighted sum of entropies obtained for each value of attribute under consideration. The weight is number of elements in the branch over the total number of elements.

Region yields maximum gain ratio. This becomes the root node of ID3 decision tree, as in Figure 48 below:

![Root Node of the ID3 decision tree](image)

Figure 48 Root Node of the ID3 decision tree

No rules are obtained as entropy of no branch is zero.

For Region = Central i.e. C, entropies are calculated for other two variables – Handicap Nature and Area. Following rules are obtained:
1. If Region = C, Area = A1, Course = BCA
2. If Region = C, Area = B2, Course = MCA

For, Region = D i.e. Delhi and other values, total of 24 rules are obtained:

3. If Region = D, Handicap = B2, Course = CIC
4. If Region = D, Handicap = C3, Course = BCA
5. If Region = E, Handicap = A1, Course = BCA
6. If Region = E, Handicap = B2, Area = A1, Course = CIC
7. If Region = E, Handicap = B2, Area = B2, Course = BCA
8. If Region = E, Handicap = C3, Course = BCA
9. If Region = E, Handicap = D4, Course = MCA
10. If Region = E, Handicap = E5, Course = BCA
11. If Region = N, Handicap = B2, Area = A1, Course = BCA
12. If Region = N, Handicap = B2, Area = B2, Course = BCA, CIT
13. If Region = N, Handicap = C3, Course = BCA
14. If Region = N, Handicap = D4, Course = BCA
15. If Region = N, Handicap = E5, Area = A1, Course = BCA, MCA
16. If Region = N, Handicap = E5, Area = B2, Course = BCA, MCA, CIT
17. If Region = S, Handicap = A1, Course = BCA
18. If Region = S, Handicap = B2, Area = A1, Course = BCA, MCA
19. If Region = S, Handicap = B2, Area = B2, Course = MCA
20. If Region = S, Handicap = C3, Course = BCA
21. If Region = S, Handicap = D4, Area = B2, Course = BCA, MCA
22. If Region = S, Handicap = D4, Area = C3, Course = BCA
23. If Region = S, Handicap = E5, Area = A1, Course = BCA, MCA
24. If Region = S, Handicap = E5, Area = B2, Course = MCA
Decision tree finally obtained is shown in Figure 49. The rules above and clusters or sets in this figure are the recommendations generated using ID3. These rules can be read from this visual itself by traversing it in top to bottom fashion starting from first node, moving along its branches to other nodes. Nodes represent attributes and branches emerging from them are their values.

Challenges and Limitations:

- **Sparsity**: The provided database was scanty and scattered.
- **Cold start**: New items and new users pose a significant challenge.

The above can be overcome by:

- Better data collection methodology
- Change in enrolment methods and forms: online, adaptive and accessible profiles and portfolios.
5.8 Discussion

Results can be used for various purposes. These categorized students can now be informed about others who are in their group. After obtaining their consent, their e-mail ids and other information can be given to their possible peers in the same group. They can be similarly grouped on the university’s portal where they do activities with a similar peer. This includes learning from others, sharing accessible notes (or material), media, bookmarks, books or e-Books (magazines, articles), places visited, programs or workshops attended. They can also share information about various schemes or benefits they are getting from Government to spread awareness. Learning in groups can be motivating and joyous.

![Diagram: Social learning (learning by observing role models)]

In the above analysis attributes like age, gender and marital status are discarded to avoid touching personal issues. Also, controversial attributes are removed like caste, religion etc. Clustering or classifying students on the basis of such attributes may not be good for their educational development.

Inclusion of attributes like Area, Region and Handicap Nature is done so that students can communicate even face to face. Their problems are likely to be same when they are from...
same Area and Region, and suffer same type of disability. This also gives a platform for context based mobile learning. This selection is demographic in nature and in accordance with mandate of the university. During census, Indian Government has collected more and more data about citizens of India and is now also assigning Unique Identification Number (UIDs) to them [102]. Categorization of subjects can be based on profiles and targets (again part of profile).

Such analysis can extended to any population for categorizing students or in general, people. These groups obtained in the rules shown are just suggestions (humanistic approach). Groups keep getting made and broken, ultimately students settle down in some company. These are recommendations to give a start to disabled students of IGNOU who joined SOCIS in 2009. Suggesting peers can initiate the process of GCI (Group Controlled Instructions) where students can discuss code, projects and work on data (if possible) as per their need and comfort. Students can learn to take up social roles [103]. Such an attempt maintains the inclusive nature of education and at the same time provides benefits of segregation as well. This chapter achieves a part of the objectives – starting a (PRS) Peer Recommender System. When this study is extended or generalized to all the students then content for all students, can be build up and better peer to peer learning environment can be created and implemented more effectively. This will build up a spirit of cooperation amongst pupil and thereby eliminating problems like ragging. Online learning and peer to peer learning can also help in decreasing dependence on coaching institutions and tuitions [92]. Moreover, university and NGOs can get in touch with students to provide links and information as per the type of group (see rules).

5.9 Summary

In this chapter, we mainly discuss classification, decision trees and state-of-art technique ID3. The discussion is followed by data analysis, results and discussion of the
results so obtained Grouping students can help in interactions. Forming of socio economic groups can help in effective peer-to-peer learning. By combining this case study with assessment data, better holistic image may come up. Moreover, this study has the repercussions to general public as it can be extended to their benefit.