Outlier Analysis of Finding Suitable Courses Based on Students Learning Styles

LakshmiSreenivasaReddy.D[1]
Department of CSE
RISE Gandhi Group of Institutions
Dr.B.RaveendraBabu[2]

Director
Delta S/w,Technologies,

Dr. A. Govardhan[3] Director of Evaluation JNTUH
Hyderabad

Abstract

Detection of outliers is an important for many tasks. The outlier detection is to find small groups of objects those do not comply with other objects. Most of the educational data consist of categorical data. In this paper we formally define the problem of outlier detection of categorical data by entropy measurement. By this approach we provide a model to guide the students to select suitable course based on his / her abilities and learning styles.

Key word:

Outlier, entropy, education data mining, ILS Questionnaire

1.Introduction

The previous model used numerical models to identify outliers like k-NN, density distance based, cluster based method etc. to identify learners' learning styles. To present customaries environment to the user to select suitable course it is necessary to create classification model on learners learning behaviors collected form ILS Questionnaire which was developed by Felder and Silverman consist of 44 questions divided in to four dimensions- active /reflective, global / sequential, sensing / intuitive, visual / auditor. Each dimension consistsof 11 questions. Along with this questionnaire we added some other personalized questions to consider factors related to his / her education environment. Many factors reduce the reliability of learners' model. Preprocessing is needed to extract the meaning full information from the collected learners' data.

2. Outliers in Education data:

2.1 What is an outlier?

Outliers are objects which do not correspond to an ideal model of the data. Outliers have extraordinary behavior comparing with other data objects. An object that is significantly different from other objects is called as outlier. These outliers lead us to wrong decision making. Outliers arise from different factors like measurement errors, data entry mistakes, giving wrong information etc.Suppose we have a system which collects learners learning information from their problem solving processand creates a learner model based on the rate of incorrect answers and the time required per problem. Sometimes the students want to finish the exam early then we can't get expected data.If we consider this data,the model would be degraded.It is needed to consider such data points to be outliers.

3.Approach

When we collect the data from students by ILS questionnaire with other personal data, the student may give wrong answers due to some factors mentioned below.

3.1.Personal problems

Some students do not want to reveal their data .For example in our collection of data there are some incomplete answers about father and mother qualificationsand their occupation,percentage of marks,brought-up profile particularly when the students live in rural area.

3.2.Psychological factor

There may be some Psychological changes in learner while attempting questions for answers.The Psychological changes can be further divided into intentional case and un-intentional case.In intentional case the student may lose his motivation to learn and responds with good regard for the intended learning result.It leads to wrong data.To rectify this data by conducting the same questionnaire to the same student more timesand by offering gifts to the students who gives the correct answers.In un-intentional case the learner data shows shifts over time as natural results of increasing or decreasing skill levels or real changes in the learning styles.

3.3.Changes in learningenvironment

Time constraint is one of the external factor and the lack of user input devices to learn, have large impact on the learning process. At this stage the student data is no longer suitable for modeling.The outlier analysis is useful to identify the environment factors.We should eliminate these outliers to get reliability of modeling. This outlier analysis is useful to find the students gradual shifting over time for dynamic learning model.As it makes the model for a new learnerwhose behavior can’t be predicted.

4.A concept for identifying suitable course by learning styles

In order to make our approach applicable, we need to know about different styles of learning

4.1. Active/Reflective learning styles

With the help of learned material, active learners who are characterized as learners preferred to process information actively. For example discussing explaining or testing it. Here the same work was done by the reflective learners alone and prefer to think about the material. The students’ preference for active learners and reflective learners indicates by communication tools like discussion forums regarding discussion and explaining doubts and explaining something more after expected by active learners. Whereas frequently reading was done by reflective learners carefully and preferred to participate passively. The self assessment tests and more exercises as well as spend overall more time on exercises were performed by active learners due to the preference of testing and trying things out. Since they preferred doing something by themselves, they are supposed to spend only time on examples rather than how the problems were solved by someone. Spending more time on reading material like content objects as well as staying larger at outlines was expected by reflective learners who likes to think and reflect about the material. The results were reflected by their self assessment tests as well as on the result pages of self assessments and exercises. Then the same question expected the answer the reflective learners in a self assessment test.

4.2.Sensing/Intuitive Dimension

The performed learning style, analyzing the performance on questions about facts and on theories, their underlying meaning which called as abstract material was performed by intuited learners. Sensing learners prefer examples by learning concrete material. The objects were learned by the intuited learner and use these examples as supplementary material. Therefore the number and time spent on content object tend to be higher and the number and time spent on examples tend to be lower. A higher interest in examples, learning existing approaches and self assessment tests, exercises to check the acquired knowledge etc. Problems were solved by sensing learners, where as intuitive learner tend to be more creative and like challenges. Therefore the better answers to questions to develop new solutions were expected by them, which required the understanding of underlying theories and concepts. The work was done by carefully and slowly with more details by sensing learners. So the self assessment tests were considered as pattern because of long or more time. The answers were carefully checked by the students before submitting. The sensing learners were expected more time on reviewing their results. Therefore the performance on questions about details can be indicated by careful details.

4.3 Visual/Verbal Dimension

Visual learners can learn very clearly by graphics, images and flow charts. Here verbal learners preferred to learn from words. So the other patterns can act by the performance on questions about graphics as well as on text. Furthermore communicating and discussing with others were liked by verbal learners. Thus the verbal learning style was indicated by a high number of visits and postings as well as high amount of time spent in a discussion forum. Visiting, reading materials such as content object more often done by verbal learners.

4.4 Sequential /Global Dimension

Sequential learners were more comfortable with details but global learners feel good in seeing the" big picture" and connections to other fields. The overviews of concepts were dealt by the performance of questions are the details serve as pattern for this dimensions by the connection between concepts and questions. The "big picture" was taken by the global learners on their interests, here outline of the course and the chapters are especially important for them. The global learning styles to interpret predefined solutions and to develop new solutions require connective topics to each other. Thus the respective questions were performed better. The navigation of learners in course acts also as a pattern denoting global learning style. The course was go throw step by step by the sequential learners in a linear way. Global learners tend to learn in large leaps, by skipping leaning objects and jumping to more complex material .So a pattern can act by the number of skipped learning objects.

5. The model

In this section we present the concept of entropy and outliers and formulate the problems.

5.1.Entropy

Entropy [6]is a measurement of information and uncertainty of a random variable (attribute).Let us take a random variable 'A' and 'S(A)' is the set of values in A, P(x) is the probability function of''A' The entropy of 'A' is defined as

$$\text{Entropy } (A) = -\sum_{V \,\varepsilon\, s(A)} P(v).\log(p(v)) \quad \ldots\ldots\ldots 1$$

The entropy of a multivariate vector $X=(v_1,v_2,\ldots\ldots..v_m)$ can be computed as

$$\text{Entropy } (X) = -\sum_{V_1 \varepsilon\, s(A_1)} \text{---} \sum_{Vm \varepsilon s(Am)} P(v_{1..}.v_m) \log p(v_1\ldots.v_m))\ldots 2$$

In education data all attributes are independent to each other. So the joint probability of combined attribute becomes the product of the probabilities of each other hence

$$\text{Entropy } (X) = -\sum_{V_1 \varepsilon\, s(A_1)} \text{----} \sum_{Vm\, \varepsilon\, s(Am)} P(v_1\ldots.v_m) \log (p(v_{1\ldots\ldots\ldots}v_m))$$

Becomes as

$$\text{Entropy } (X) = \text{Entropy } (A_1) + \text{Entropy } (A_2)\text{---}$$

$$+\text{Entropy } (A_m)\ldots\ldots\ldots\ldots.. 3$$

Problem formulation:

Given a dataset D of npoints X_1 X_2 - - - -$X_n$where each point is a multidimensional vector of m categorical attributes i.eX_i $=(v_{1i},v_{2i},\ldots\ldots\ldots\ldots\ldots\ldots\ldots.v_{mi})$,

where V_{1i}Єs(A_1)……….v $_{mi}$Є s(A_m) and |D|=n. |x_i|=n

And given an integer 'k' the no of outliers to be found and let the set of outliers as OCD

Where |D|=n ,|D-O|= n-k such that the

Entropy of (D-O) isminimum i.e. min Entropy (D –O) subject to |D|=,k O ε D

We repeat to find Entropy of (D-O) for maximum no of $|D|_{Ck}$times comparing with previous entropy value. If the Entropy of i^{th} step combination of records is less than j^{th}step combination of recordswe retain that i^{th} combination and leave the j^{th}combination of records. Next we select another combination of records in j^{th}place .otherwise we replacei^{th}combination. By recursively comparing Entropy values in each step we can find the k outliers such that (D-O) gives minimum entropy.

5.Experimental Results

We have applied the preceding method on 256 engineering students of ECE,EEE,CSE branches in RISE Gandhi and Rise Prakasam Institutions.We have collected data from ILS questionnaire which consists of 44 questions of four different learning styles developed by FELDER and SILVERMAN,and 10 other attributes related to personal environment and 1 class label attribute which consists of class labels

ECE,CSE,EEE.We have classified the data by decision tree method of using 60% of the data as training data set and 40% as test data set.

After eliminating the 3 outliers by proposed model and classifying using decision tree method of using 60% of data as training data set and 40% of test data from the remaining 253 data objects/records the error rate has minimized comparing with original data.We repeated it 5 cross validation trails.

6.Conclusion

We are developing a system that can guide the students in choosing the suitable branch(course)in Engineering based onhis/her learning style to succeeded in their field.We use a decision tree machine learning technique to build this model and present a outlier elimination using entropy approach. Experimental result indicates that outlier elimination improves the reliability of the decision tree.In future work we will further analysis how changes in proposed model effect the improvement in the data.

References

[1] Han J kamsserM ,“Data mining concepts and techniques elseiever”,2001

[2] Tal Bok yorn,“IEEE work shop improvement of learning styles diagnosis based on outliers reduction of user interface behaviors’,2007

[3] yongsekim“IEEE cont outlier analysis of learners data based on user interface behaviors”,2007

[4]Silvia Rita viola “Analysis of Felder-Silverman index of learning styles by a data driven statistical approaches”,2006

[5] Zengyou He “An optimization model for outlier detection in categorical data”

[6] C.E.Shannon “AMathematical theory of ommunication.Bell System Technical Journal”,1948,pp.379-423