International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Time Variant Multi Perspective Hierarchical Clustering Algorithm for Predicting Student Interest in Sports Mining

A.Basheer Ahamed, M. Mohamed Surputheen

Abstract: ***Predicting performance of students in sports is analyzed and studied. There are many techniques identified for the prediction of sports interest and they are not producing expected value. Towards performance development, a novel time variant multi perspective hierarchical clustering approach towards user interest prediction. The proposed time variant model reads the sports log and groups them according to the time domain. The entire log has been split into different of clusters as like time window. Then using window log, the method splits the logs according to different sports. For each time window, the method identifies the list of actions or sports played or tagged or chat with other users. Using the class of log, the method identifies the category of sports log and for each category of sports, the method compute the sports strike strength (SSS). Based on the value of SSS, the method identifies the user interest. Similarly, the interest of the student at each time window has been identified and used to generate the knowledge. The proposed method improves the performance of sports interest prediction on students with less false ratio.***

Keywords: Sports Analysis, Data Mining, ML, Sport Prediction, Hierarchical Clustering, TMHC, Sports Mining.

I. INTRODUCTION

The modern society has been diverted from spending time for sports and the ratio of playing games in student groups are getting down every year. As the students engaged in various activities like watching Tv, surfing internet, chatting time in social media, their activity in sports are getting reduced every year. This really affects the national performance in sports activity and the national sports community faces challenges in identifying the sports person with good qualities. So by analyzing the activities of students in their sports involvement, the detection of exact sports person can be identified.

The sports analysis is the process of analyzing the activity of students of any organization. Any student would be spending some time in playing games or they would chat with the friends of social media about the game. Such data has been collected and stored in sports data set which contains various information related to the student of any institution, which combines, personal, educational, family, social group, their chats and so on. By analyzing the data, you can identify the interest of the user in specific or multiple sport. The result of such analysis can be used to generate useful knowledge to the decision makers.

The analysis of sports data can be handled using different data mining techniques. As like the data mining algorithms has been applied to several problems, the sports data analysis can also be performed using data mining techniques. The student would be having multiple interests in various domains. As for as the sports data considered, the same student would be having different sports interest. So in order to predict the user interest, it is necessary to group them. Further the sports data can be classified in number of levels or hierarchy. For example, the “cricket” data can be classified as playing cricket, commendatory, umpiring, and so on. So in order to produce efficient clustering performance, the sports data should be classified in multiple levels. The k means algorithm cluster the data according to the distance measure which is scalable in dimension. The fuzzy c means algorithm uses only range values which introduce poor clustering. Similarly, the support vector machine (svm) clustering introduces overlap in clusters. The decision tree based approaches makes decision only based on the features of nodes and restricted on number of dimensions. Similarly, there are many approaches available which consider only the limited number of dimensions or features in making decision. This introduces higher false ratio and reduces the performance of sports prediction.

Sports mining are the process of mining information or knowledge from sports data related to the expectation. For example, the knowledge which support decision maker is to be generated from sports data. To perform this, the list of students who are more active in specific sport can be generated. Also, the list of users who are interested in sports also can be generated. But the second one is a generalized one where the first one is precise one. In order to make decisions, the result should be precise one. In the second result, the user would be having multiple interests but he may not be specialized and precise in playing the sport considered. Such tuples are not support the growth of achievements.

Manuscript published on November 30, 2019.
* Correspondence Author
A.Basheer Ahamed*, Research Scholar, Department of Computer Science, Jamal Mohamed College (Autonomous), (Affiliated to Bharathidasan University), Tiruchirappalli, Tamilnadu, India.
Dr. M. Mohamed Surputheen, Associate Professor, Department of Computer Science, Jamal Mohamed College (Autonomous), (Affiliated to Bharathidasan University) , Tiruchirappalli, Tamilnadu, India.

Sports prediction is the process of identifying the user interest in specific or multiple sports. Based on the list of sports interested by the student, a single entity interest which is persistent must be identified. To analyze and predict the sports interest, it is necessary to consider the user interest in different time window. By considering all this, the paper presents a time variant multi perspective hierarchical clustering and Interest Prediction algorithm.

II. RELATED WORKS

Different algorithms and methods are discussed earlier for the prediction of sports interest and analysis of sports data. Such methods have been discussed in this section.

In [1], the author performs a analysis on school failure using data mining technique. The prediction model uses the logs of students belong to 600 plus middle level schools in Spain. The method uses the decision trees and induction rules. In [2], the author presents a survey on web based education systems. The quality of e-learning systems has been analyzed using different data mining algorithm.

In [3], an educational data mining technique has been used to evaluate the performance of educational institutions. The paper surveys different institutions according to the user group, environment and the kind of data given. In [4], the student data analysis is performed using a behavior analysis model (DTree-BAM). The method analyses the behavior of students and index the logs in the decision tree. The same has been used to perform classification and prediction. The method is focused to identify the willing of students in continuing their education.

In [5], the performance of secondary school student has been analyzed and predicted using different algorithms like SVM, RF and NN. In [6], the techniques of data mining are applied for the analysis of student academic failure. Different data mining approaches are used for the classification of students and their state of art. In [7], an efficient mining model is presented to identify the interest of user in different subjects and how the user interest can be predicted. By predicting the subject of interest, the performance of teaching model can be measured and analyzed.

In [8], an regression tree based success prediction model is presented towards student performance analysis. In [9], the author present a survey on level set methods towards classifying and comparing the schemes discussed in literature towards performance analysis of students. In [10], the author present Euler Clustering algorithm which uses Stiegel-manifold-based gradient method. The Euler K means algorithm produces higher performance in clustering the student records. In [11], a multi objective firefly algorithm (MFF-GSO) is presented to optimize the clustering performance. The method uses firefly algorithm with Group Search Optimizer (GSO). The objective functions are used to measure the fitness value like fuzzy DB index.

In [12], the author presents a classifier and prediction algorithm to support telecom industry. The analysis is performed with UCI repository provided by Customer Churn Dataset. In [13], a predictive modeling has been presented for the student performance measurement. The method clusters the student according to their answers. The clustering is performed using K means. The student behavior has been predicted based on the average assistance score estimated based on the student answers.

In [14], the student extracurricular activities are used to predict the performance od students using decision trees. The academic performance has been measured according to the extracurricular activities of students. By analyzing the extracurricular activities the involvement in sports is implicitly monitored. In [15], an Ant colony based sports performance has been analyzed. The similarity rules has been used to classify the data into different types. The clustering is performed using k means with improved one combined with Ant colony (ACO-Kmeans) has been proposed.

In [16], an decision tree induction approach called Multivariate Regression prediction model M5P. The method is developed to predict the student performance according to the efficiency in problem solving, learning skill, adaptability, participation in sports, amount of time spent on social networks and so on.

All the methods suffer to achieve higher performance in the prediction of student performance.

III. TIME VARIANT MULTI PERSPECTIVE HIERARCHICAL CLUSTERING AND INTEREST PREDICTION

The proposed time variant multi perspective model reads the input sports data. Retrieved data set has been preprocessed to eliminate the noisy records. The noise removed data has been split into number of sub set according to the log time in different time window. At each window log, the features are extracted and multi factor support weight for different sports class is measured to cluster the data. Similarly, the method estimate sports strike strength to predict the user interest.

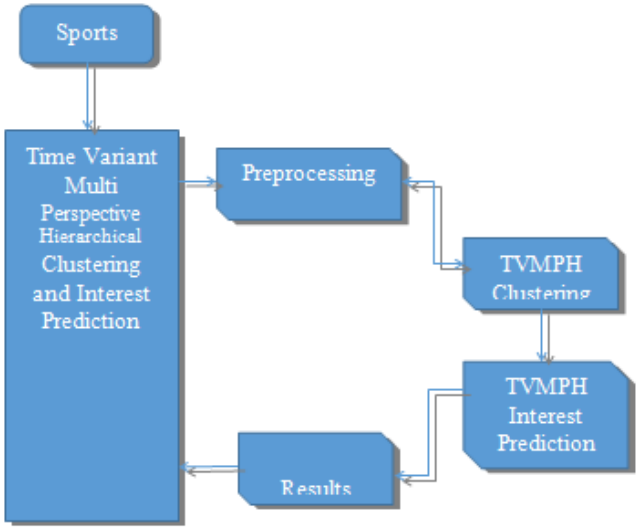

Fig. 1. Architecture of TVMPH Sports Analysis and Interest Prediction System

A. Preprocessing

The preprocessing on sports data has been performed to eliminate the noisy records present in the data set. To perform this, the entire data is read and identifies the list of dimensions or features of the log throughout the data set.

Now, with the list of features or dimensions identified the method reads each data point from the data set and verifies the presence of all the feature and values. For each data point, the presences of all the dimensions are verified and their values also verified and fill or remove invalid values. If any of the data point is identified as incomplete then it has been concluded that noisy and such records are eliminated. The preprocessed data points are used to perform clustering in the next stage.

TVMPH-Preprocessing Algorithm:
Given: Sports Data Set Sds
Obtain: Preprocessed Data Set Pds
Start
Read sports data set Sds.
Initialize Dimensions list
Dl = $\int_{i=1}^{size(Sds)} \int_{j=1}^{size(Sds(i))} \sum(Dimension \in Dl) \cup Sdsij \nexists Dl$
For each sports log sl
If $sl \in (\forall Dimesion \in Dl)$ then
Add to preprocessed dataset Pds.
Pds = $\sum(sl \in Pds) \cup sl$
End
Stop

The preprocessing algorithm takes input data given and identifies the list of features or dimensions available. With this information, the presence of all the features in each log has been verified and noisy data points are eliminated from sports data set.

Time Variant Multi perspective hierarchical clustering (TVMPH)

The clustering is performed with the preprocessed data set. The preprocessed log has been split into different sub set based on the time window when the log is generated. The user would have different interest in different time window and would be changing all the time. Also, it is necessary to find the interest of user in different time space. So it is necessary to cluster the log under different subspace. For each time window log, the method initializes number of clusters according to the number of sports interest considered. For each class of interest, the method computes frequency measure according to the total logs available and the number of plays or visits of the class and the total number of interests. Using all these information, the method computes the Multi Factor Sports Weight. Based on the value of multi factor sports weight, the method selects a class to index the log.

TVMPH Clustering Algorithm:
Given: Preprocessed Data Set pds, Sports Set Ss
Obtain: Cluster Set Cls
Start
Read preprocessed data set Pds, sports set ss.
Identify the number of time window
Tw = $\int_{i=1}^{size(Pds)} \sum Timespace \ni Tw$
Now initialize the clusters Cls.
For each sports s from ss
Create cluster
c = $\int_{i=1}^{size(ss)} \int_{j=1}^{size(Tw)} Initialize(c(s, Tw(j)))$
End
For each time window Ti
For each sport s
Compute number of visits Nov.
Nov = $\int_{i=1}^{size(Pds)} \sum Pds(i) = Ti \;\&\&\; Pds(i).sport = s \;\&\&\; Pdsi.Type{=}Visit$
Compute number of plays Nop.
Nop = $\int_{i=1}^{size(Pds)} \sum Pds(i) = Ti \;\&\&\; Pds(i).sport = s \;\&\&\; Pdsi.Type{=}Play$
Compute number of Tagged Not.
NoT = $\int_{i=1}^{size(Pds)} \sum Pds(i) = Ti \;\&\&\; Pds(i).sport = s \;\&\&\; Pdsi.Type{=}Tagged$
Compute number of interest present in the time window Ti.
Ns = $\int_{i=1}^{size(Pds)} \sum Distinct(s) \in Pds$
Compute sports frequency
Sf = $\frac{\sum_{i=1}^{size(Pds)} Pds(i).Tw==Ti \;\&\&\; Pds(i).sport=s}{\sum_{i=1}^{size(Pds)} Pds(i).Tw==Ti}$
Compute multi factor sports weight MFSW.
MFSW = $\frac{(nop+nov+not)}{\sum_{i=1}^{size(Pds)} Pds(i).Tw=Ti} \times \frac{1}{Ns}$
End
Choose the sport class with maximum MFSW.
Index the logs to the selected sport class.
End
Stop

The time variant multi perspective hierarchical clustering algorithm reads the preprocessed data set. Initializes number of clusters according to different time window where each cluster has number of subspace according to the number of interest or sports class. Further the method estimates multi factor sports weight for each sports class to identify the class for the input sports log.

TVMPH Interest Prediction

In this stage, the method reads the user clusters generated. The method computes the sports strike strength (SSS) for all the clusters which represent a time window. Based on the value of SSS, the method selects single sports for the single user. Finally, based on the occurrence of sports in the identified interest from different time window, a single one has been selected as the user sports interest. Predicted interest has been used to generate recommendations and decision making.

TVMPH Interest Prediction Algorithm:
Given: Cluster set Cls, Student st
Obtain: Interest Intr.
Start
Read cluster set cls, St
Initialize interest set Ins.
Identify the number of time window
Tw = $\int_{i=1}^{size(Pds)} \sum Timespace \ni Tw$
Identify user log
Ul = $\int_{i=1}^{size(Pds)} \sum Pds(i).User == St$
For each sport s
For each time window Ti
Compute number of visits Nov.
Nov = $\int_{i=1}^{size(Ul)} \sum Ul(i) = Ti \;\&\&\; Ul(i).sport = s \;\&\&\; Ul(i).Type = Visit$

Compute number of plays Nop.

Nop = $\int_{i=1}^{size(Ul)} \sum Ul(i) = Ti \;\&\&\; Ul(i).sport = s \;\&\&\; Uli.Type=Play$

Compute number of Tagged Not.

NoT = $\int_{i=1}^{size(Ul)} \sum Ul(i) = Ti \;\&\&\; Ul(i).sport = s \;\&\&\; Uli.Type=Tagged$

Compute sport strike strength SSS.

SSS = $\frac{Nov \times \alpha}{\sum_{i=1}^{size(Ul)} Ul(i).Tw=Ti} \times \frac{Nop \times \beta}{\sum_{i=1}^{size(Ul)} Ul(i).Tw=Ti} \times \frac{NoT \times \gamma}{\sum_{i=1}^{size(Ul)} Ul(i).Tw=Ti}$ where $\alpha - 0.6\ and\ \beta - 1.5\ and\ \gamma - 0.8$

End

End

Compute cumulative SSS as

CSSS = $\frac{\sum_{i=1}^{size(Tw)} SSS}{size(Tw)}$

Find the interest with maximum CSSS value.

Interest or sports

Intr = $\int_{i=1}^{size(sports)} Max(sports(i).CSSS)$

Stop

The TVMPH interest prediction algorithm estimates the sports strike strength towards different interest at each time window. Based on the value of sports strike strength the method selects a single interest as result.

IV. RESULTS AND DISCUSSION

The proposed time variant multi perspective hierarchical clustering algorithm and interest prediction method is hard coded. Its performance is evaluated in various conditions. The method is implemented using advanced java. The performance of the method is verified under various parameters. The results obtained is presented in this section and compared to the performance of other methods.

Table- I: Evaluation Details

Parameter	Value
No of Sports	10
Number of Sub class	3
Number of Students	5000
Number of Time window logs	2 years

The parameters considered for the evaluation of the TVMPH algorithm is presented in Table 1. The evaluation is performed with the logs generated in two years of 5000 students which falls under 10 different sports class where each class has 3 levels of sub classes. The performance of the proposed algorithm is matched with others.

Fig. 2. Performance on clustering accuracy

The clustering accuracy produced by different methods is compared in Figure 2. The proposed TVMPH algorithm achieved higher accurate clusters.

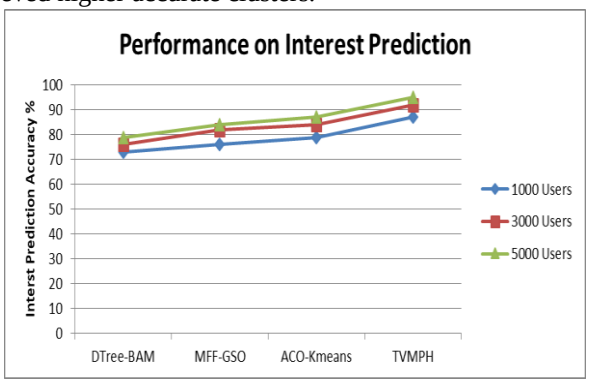

Fig. 3. Performance on Interest Prediction

The performance on interest prediction accuracy is measured and compared with others. The proposed TVMPH algorithm achieved higher interest prediction accuracy.

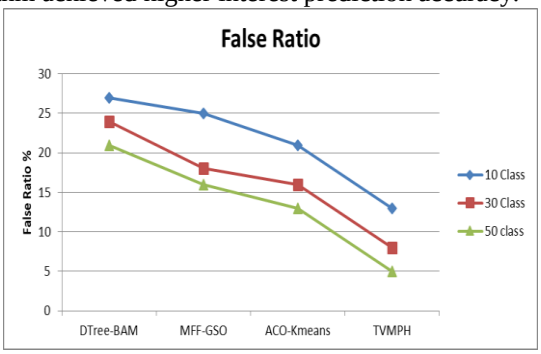

Fig. 4. Performance on false ratio

The false classification ratio produced by various methods is compared in Figure 4. The TVMPH algorithm introduced only less false ratio than other methods.

V. CONCLUSION

In this paper, an efficient time variant multi perspective hierarchical clustering algorithm has been presented. The method reads the sports log and groups them under number of classes. For each time window, the method generates number of clusters according to the number of interest or sports available.

While clustering, the estimates the multi factor sports weight towards various sports according to the number of visits, plays, and tags towards various sports class. Based on the value of multi feature sports weight the clustering is performed. On the interest prediction, the method estimates the sports strike strength (SSS) to predict the student interest. The method produces higher performance in clustering and interest prediction with less false ratio.

REFERENCES

1. Carlos Marquez-Vera, et. Al , "Predicting school failure and dropout by using data mining techniques". IEEE Journal of Latin-American Learning Technologies, Vol. 8, No. 1, February 2013.
2. C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems. Appl., vol. 33, no. 1, pp. 135-146, 2007.
3. C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," IEEE Trans.Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 6, pp. 601-618, Nov. 2010.
4. V.P. Bresfelean, "Analysis and Predictions on Students' Behaviour using Decision trees in WEKA Environment", Proceedings of the ITI 2007 29th Int Conf. on Information Technology Interfaces, June 25-28, 2007.
5. P.Cortez and A.Silva, "Using Data Mining To Predict Secondary School Student Performance", In EUROSIS, A. Brito and J. Teixeria (Eds.), 2008, pp.5-12.
6. P. Bresfelean , et. Al " Determining Students' Academic Failure Profile Founded on Data Mining Methods", Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces, June 23-26 2008.
7. B.K. Baradwaj, S. Pal, "Mining Educational Data to Analyze Students' Performance", (IJACSA) International Journal of Advanced Computer and Application, Vol. 2, No. 6, 2011.
8. Z.J.Kovacic(2010),"Early Prediction of Students Success: Mining Student Enrolment Data", paper presented at proceedings of Informing Science & IT Education Conference(InSITE), casino Italia,June,19-24-2010.
9. Saima Sayyed, Rugved Deolekar, A Survey on Novelty Detection using Level Set Methods ICICCT,.2017.
10. Jian-Sheng Wu, Wei-Shi Zheng Euler Clustering on Large-scale Dataset IEEE 2017.
11. Latha Parthiban, Multi objective hybridized firefly algorithm with group search optimization for data clustering Golda George, IEEE ,2016.
12. Ahmed, M., et. Al Mcs: Multiple classifier system to predict the churners in the telecom industry .2017.
13. Alana M. de Morais, et. Al " Monitoring Student Performance Using Data Clustering and Predictive Modelling " IEEE,2014.
14. Tomas Hasbun, et. Al "Extracurricular activities as dropout prediction factors in higher education using decision trees"IEEE 2016.
15. Jian?et. Al "Clustering Analysis of Sports Performance based on Ant Colony Algorithm"IEEE 2016.
16. [16] S Chaitanya Kumar, et. Al "M5P Model Tree in Predicting Student Performance : A Case Study" IEEE,2016.

AUTHORS PROFILE

A.Basheer Ahamed, Working as an Assistamt Professor with 11 years of experience at Jamal Mohamed College(Autonomous) (Affiliated to Bharathidasan university,) Completed PG Computer Science in Jamal Mohamed College, and M.Phil in Periyar University. Passed in TNSET Exam 2016. Guided two M.Phil Scholars. Successfully completed the course “Positive Psychology” awarded by NPTEL-AICTE Faculty Development Programme. Currently acting as an organizer for Spoken Tutorial conducted by IIT Mumbai and as a web administrator of Jamal Mohamed College (Autonomous).

Dr.M.Mohamed Surputheen, Working as a Associate Professor in Jamal Mohamed College with more than 20yrs experience. Education Qualification is MSc..Mphil and PhD. Guiding Mphil Scholars and around eight PhD Scholars. Published and Presented more than twenty five research papers. Currently acting as a Controller of Examination in Jamal Mohamed College (Autonomous) Communication Mail ID: abammb2@gmail.com.