BANDWIDTH MANAGEMENT AND MINING OF STUDENT'S BEHAVIORAL PATTERNS USING HYBRID DATA MINING TECHNIQUE

[1]R Vijaya Lakshmi, [2]Radha R, [3]Bhoomika A.P
[1]Assistant Professor, [2] Assistant Professor, [3] Assistant Professor
[1]Computer Science and Engineering,
[1]Alliance University, Bangalore, Karnataka, India

Abstract: One of the significant contests in managing multi-service computer networks is to accomplish the bandwidth effectively which plays a vital role in many internet services. In these networks, such as university networks, there is often constraint on bandwidth assets and the managers must allot it in an effective and appropriate method. The active bandwidth management in multi-service computer networks such as university networks has become a contest in current years. The progress of internet traffic and curb of bandwidth resources influence the information technology (IT) directors to focus on effective bandwidth provision policies. One of the vital matters debated in this part is how to allot the bandwidth fairly built on the priority levels. In this paper, concentrating on the "priority-based bandwidth allocation", a hybrid data mining way is advanced to accomplish the partial bandwidth in a university network extra effectively. This way is composed of two main stages and customs the bunching and organization techniques. The main resolution is to detect, analyze and foresee students' behavioral patterns in a university network and classify the main issues that touch their bent in using internet.

Keywords: Bandwidth management, Data mining, Clustering, Decision tree.

I. Introduction

Data mining is the group of exponentially increasing methods which are used to find some valuable data, patterns and information from already given data [1]. This valuable data helps to advance existing research and recover productivity. The requirements of data mining are innumerable; it is used nearly in every facet of life [2]. Some main applications of data mining are Healthcare ,Market Analysis, Finance, Education, Manufacture Engineering, Corporation Investigation and Agriculture [3]. The data mining methods can around discriminated into two kinds predictive & descriptive [4]. The predictive family can additional order into organization, regression and time series study. In this paper we limit our exploration room to apply data-mining systems for crops disease and loss prediction which goes to sorting (predictive) family. Data mining systems help in agriculture for the prohibiting of manual jobs and for decision making which enable to decrease manufacture cost and recovers efficiency [6].

We essence on assembly and classification techniques in this paper and a new hybrid method is progressive. Primarily, all of the students are collected based on their social features in intense internet by the clustering technique. In this stage, the results of clustering are inspected and dissimilar groups of students are graded based on their behavior in spending internet. Positioned on the clustering results, as new variable is clear indicating the behavioral patterns of students in spending internet. Then, the logistic regression and decision tree techniques are used in directive to find the factors that touch students' behavior in spending internet. This method can also be obliging in order to foresee students' tendency in consuming internet based on their characteristics.

II. Background and Related work

In this unit, primarily the classification methodology is deliberated which aids to know this paper more easily. Then the present linked methods will be discussed.

A. Classification

Classification is one of the famous data-mining subfields that assign information in a group for directing predefined nomenclatures or classes. Classification has straightforward objective to properly predict the target class for to each data record. For instance, such model (classification) can be used to forecast crops damage as low or high.

B. Clustering and K-means algorithm

Clustering is an unsupervised method and tries to split data into approximately groups such that the objects in to each cluster are very identical while being dissimilar to the data from other clusters [21], [20], [19]. Clustering algorithms can be categorized into two major groups: hierarchical and partitioning. The K-means algorithm is one of the isolating clustering methods and has been extensively used for clustering. The correctness of this algorithm depends on the primary centers [14], [15]. It usually uses the Euclidean distance and needs the number of clusters (K) as input [22].

At initial, this algorithm receipts K as input and arbitrarily selects K of the items as the primary centers of the clusters. At the second step, the left over objects are assigned to the closest clusters grounded on the distance between the object and the center of

the cluster. After that, it computes the average of the items as the new center for each cluster. This process repeats until the standard function converges

C. Logistic regression

The logistic regression is a procedure of multiple regression models with a categorical reliant on variable. This categorical variable can acquire two or further values. It is important that the predictor variables could be a grouping of continuous and categorical fields [21 and 12]. The logistic regression computes the likelihood that a particular object with exceptional values falls in one of the categories of the reliant variable. The common linear regression model is shown in formula (1) which p indicates the probability stated above.

Ln(p/1-p)= _0 + _1X1 + _2X2 + … + _kXk

(1) shows the constants of the regression model and Xi specifies the predictor variables [11].

Decision tree

Decision tree learning approaches use branching method to demonstrate every possible outcome of a decision. They can effort with discrete-value attributes and continuous value attributes as well. The learned trees are then characterized in the form of if-then rules. Three elementary elements of the tree are decision node, branch and leaf node. Decision tree can be also used for attribute subset selection and definition the best features. In this case, each attributes that do not appear in the tree are measured as irrelevant features; the attributes appearing in the tree form the nominated subset of features and the most significant attributes. The output of decision trees can be transformed to classification rules [37 and 43].

Implementation

In this paper, a hybrid data mining technique is proposed and applied on an actual dataset collected from a university. The data and the method are individually discussed in the following sub sections.

Analysis of data

As mentioned earlier, the data belongs to a university network. This data covers the detailed information about each student and his/her conduct in using internet. Actually, it helps the IT managers of the university to screen the number of times that each student has used the internet for diverse purposes such as scientific, entertainment, general or checking the email. The features of each student are also provided. An example of the data is shown in Table 1. The first six variables demonstration some of the demographic features of the students. The next variable (FS) specifies the sum of times that the student visited the scientific web sites. The two following fields (FE and FG) express the frequency of visiting the web sites of entertainment and general, correspondingly. Finally, the last variable (FC) is the number of periods that the student used internet to check email. The total number of records is 12709 which compact to 11609 after preprocessing. Supervising the noise and missing values were considered in the data preprocessing phase.

III. Proposed Method

As mentioned earlier, a new hybrid data mining method is offered to detect, analyze and predict the students' behavioral patterns of using internet in a university network. This method comprises two main phases: "Mining students' behavioral patterns in using internet" and "Mining the factors that disturb students' behavioral patterns in using internet". These two phases are discussed in details as follows.

1) Mining students' behavioral patterns in using internet

In this phase, a grouping technique is used to extract different groups of students with various propensities in using internet. To achieve this objective, the students are clustered grounded on their behavioral features. The variables are the last four columns of Table 1 shown as FS, FE, FG and FC. These variables are worthy measures to indicate the students' conduct and tendency in using internet based on the opinions of IT managers of this university network. They established an automotive system which events these variables. The K means algorithm is applied to cluster the students based on these four fields; the sum squared error (SSE) measure is used to make out the optimum number of clusters. The chief purpose of grouping the students based on the four variables is to classify different behavioral patterns by examining the results of clustering. For example some of the students may use internet typically for positive and good drives such as visiting scientific sites while others have inclination to use internet only for entertainment or checking their emails. In this way, one of the goals of clustering is to classify the group of students who use internet in a positive way, like the first stated group. So, by analyzing the results of clustering, different behavioral patterns in using internet are recognized. Based on mined patterns, a new variable is defined that indicates the type of student's tendency in consuming internet. Accordingly, this stage is composed of three steps as shown in Figure 1. In the following phase, using the extracted new field as dependent variable, the logistic regression and decision tree techniques are applied to predict students' tendency to use internet for scientific purposes.

2) Mining the factors that affect students' behavioral patterns in using internet

In this point, considering the fallouts of clustering, the logistic regression and decision tree techniques are used to find the factors that affect the style of using internet by students. The new field defined based on the consequences of clustering, is entered in the model as dependent variable and student's features are used as independent variables. The independent variables comprise Gender, Age, Program, Degree level and Average score.

Student number	**Age**	**Gender**	**Program**
124	21	M	UG
125	22	F	UG
Student number	**Degree level**	**Average Score**	**FS**
124	MS	16.45	52
125	MS	16.12	54
Student number	**FE**	**FG**	**FC**
124	125	121	48
125	13	111	28

Figure: 1

Phase 1- cluster analysis: drawing out the students' behavioral patterns in using internet
Phase 2- logistic regression & decision tree modeling: drawing out the factors that affect students' behavioral patterns in using internet and the estimate of students' tendency
Phase 3- managerial analysis: The aim of this phase is to test the relationship between different characteristics of students and their tendency to use internet in an optimistic way like scientific purposes.

The effect and impact degree of each variable could be deliberated in this step by inferring the results. This mined model could be also useful for predicting purposes. Surely, the IT managers are attentive to be able to predict a new student's tendency in using internet given the student's age, gender and other features. In fact, the ability to forecast that a student would use internet in an optimistic way like scientific purposes, could be very supportive. So, this phase is poised of two steps as shown in Figure 1: detecting the significant factors and the guess of students' tendency.

3) Managerial analysis

The outcomes are analyzed in this phase. The chief purpose is to apply the results and mined knowledge in order to a healthier bandwidth management and allocation. In this phase, we try to make proposals to the IT managers of the university rendering to the results of the two previous phases. In other words, we discuss on how the IT managers can use these consequences in order to a better bandwidth management. The mined behavioral patterns and the main factors that have influence on student's propensity are analyzed. Finally based on the analysis, managerial implications are presented for the IT managers of the university.

IV. Results

The results of the projected method are offered in this section. The results are deliberated step by step as shown in Figure 1.
Step 1: In this stage, all of the students were grouped in terms of FS, FE, FG and FC. These four variables were normalized by the min-max normalization method. The K means algorithm was employed in order to cluster the students and the SSE measure was used to find the optimum number of clusters. The value of this index for dissimilar values of K was computed which is shown in Table 2. Based on the results, when K variates to 3, the value of the SSE is minimized and the variations are smooth. So, the initial number of clusters was set to 3. The schema of clusters considering FS and FG is shown in Figure 2.
Step 2: In this step, the clusters are understood seeing the values of the four variables. The centers of individual cluster are shown in Table 3. The centers of clusters are the average standards of fields.

V. Results Analysis and Managerial Implications

In this segment, we discuss on how the results of phase 1 and phase 2 could be supportive to better bandwidth management. According to the results, there are two key behavioral patterns in using internet in this university; the students who use internet for scientific purposes and those who use it for non-scientific purposes. Accordingly, the IT managers can clarify two priority levels and allocate the bandwidth accordingly in the state they face limited bandwidth recourses.

Furthermore, the main factors which effect the students' tendency in using internet for scientific purposes were recognized as "Age" and "Degree level". Surely, the IT managers are interested to expect that a student would have tendency to use internet for scientific purposes or not. This prediction for a new student could be very useful in nominal bandwidth allocation. In brief, the results help the IT managers to notice the groups of students who use internet for scientific purposes to get the high priority to these students.

VI. Conclusion and Future Research

Bandwidth management is one of the vital issues in university networks which many scientists have focused on. In this paper, directing on the "priority-based bandwidth allocation" in bandwidth management, a hybrid data mining method based on the

clustering and decision tree was offered and applied on a real data of a university network. The key result is that this method is capable in this area. The results show that the projected method could provide the IT managers with valuable knowledge to develop their decisions. The results of clustering designate that different groups of students are alike in using internet for entertainment drives and checking email; in fact the frequency of visiting scientific web sites and the general ones (FS and FG) are the fields that can label different behavioral patterns in using internet. Accordingly, the ratio of these two variables can be used to distinguish the affinity to use internet by students. Furthermore, two major groups of students are well-defined as the scientific and nonscientific students. The grades of implementing the decision tree display that the student's "Age" and "Degree level" are the main factors with the weights 0.52 and 0.422, respectively. The "Gender" and "Average score" aspects have the same weights equals 0.0285. The weight of the trait "Program" is zero. Accordingly, the "Age" and "Degree level" are recognized as the most prominent factors that affect students' tendency in using internet. Furthermore, based on the results, computing the probability that a student would have affinity to use internet for scientific purposes is possible based on his/her characteristics. Added research is needed to improve the bandwidth management; specially using the operation research (OR) models to enhance the bandwidth allocation. At the next step, we target to combine the proposed method with OR models for more active bandwidth allocation among diverse groups of students. By assimilating the proposed method in this paper and OR models, we can specify the quantity of bandwidth for each student more effectively based on his/her tendency to use internet.

References

[1] S. K. Nair, and D. C. Novak, "A traffic shaping model for optimizing network operations", *European Journal of Operational Research*, Vol. 180, pp. 1358–1380, August 2007.
[2] S. Deb, A. Ganesh, and P. Key, "Resource allocation between persistent and transient flows", *IEEE/ACM Transactions on Networking*, Vol. 13, pp. 302 – 315, April 2005.
[3] N. Ni, and L. N. Bhuyan, "Fair Scheduling in Internet Routers", *IEEE/ACM Transactions on Networking*, Vol. 13, pp. 686-701, June 2002.
[4] W. Ogryczak, M. Milewski, and A. Wierzbicki, "On fair and efficient bandwidth allocation by the multiple target approach", *IEEE Next Generation Internet Design and Engineering*, pp. 8-55, April 2006.
[5] Z. Deyu, and L. Zhiguo, "Research on Cluster-Based Bandwidth Allocation Algorithm in Ad Hoc Network", *IEEE Hybrid Intelligent Systems*, pp. 268 -270, August 2009.
[6] D. Kumar, K. Murugesan, S. Raghavan, and M. Suganthi, "Neural Network based Scheduling Algorithm for WiMAX with improved QoS Constraints", *IEEE Emerging Trends in Electrical and Computer Technology*, pp. 1076– 1081, March 2011.
[7] B. R. Chang, and H. F. Tsai, "Improving network traffic analysis by foreseeing data- acket-flow with hybrid fuzzybased model prediction", *Expert Systems with Applications*, Vol. 36, pp. 6960–6965, April 2009.
[8] T. Anjali, C. Bruni, D. Iacoviello, G. Koch, and C. Scoglio, "Filtering and forecasting problems for aggregate traffic in Internet links", *Performance Evaluation*, Vol. 58, pp. 25–42, October 2004.
[9] V. S. Frost, and B. Melamed, "Traffic modeling for telecommunications networks", *IEEE Communications Magazine*, Vol. 32, pp. 70–81, March 1994.
[10] M. Krunz, and H. Hughes, "A traffic model for MPEG-coded VBR streams", *ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, Vol. 23, pp. 47–55, May 1995.
[11] D. P. Heyman, and T. V. Lakshman, "Source models for VBR broadcast-video traffic", *IEEE/ACM Transactions on Networking*, Vol. 4, pp. 40–48, February 1996.
[12] P. Bocheck, and S. F. Chang, "A content based video traffic model using camera operations", *IEEE Image Processing*, Vol. 1, pp. 817-820, September 1996.
[13] P. R. Chang, and J. T. Hu, "Optimal nonlinear adaptive prediction and modeling of MPEG video in ATM networks using pipelined recurrent neural networks", *IEEE Journal on Selected Areas in Communications*, Vol. 15, pp. 1087–1100, August 1997.
[14] D. Doulamis, N. D. Doulamis, and S. D. Kollias, "Nonlinear traffic modeling of VBR MPEG-2 video sources", *IEEE International Conference in Multimedia and Expo*, Vol. 3, pp. 1318–1321, August 2000.
[15] Bhattacharya, A.G. Parlos, and A.F. Atiya, "Prediction of MPEG-Coded Video Source Traffic Using Recurrent Neural Networks", *IEEE Transactions on signal processing*, Vol. 51, pp. 2177-2190, August 2003.
[16] W. Junsong, W. Jiukun, Z. Maohua, and W. Junjie, "Prediction of Internet Traffic Based on Elman Neural Network", *IEEE Chinese Control and Decision Conference*, pp. 1248-1252, June 2009.
[17] F. Pilka, and M. Oravec, "A NARX neural network algorithm for video traffic prediction", *Journal of Electrical and Electronics Engineering*, Vol. 4, pp. 179-184, May 2011.
[18] G. Moayeripour, A. Aghakhani, M. N. Moghadam, and H. Taheri, "Reducing bandwidth allocation delay in a DVBRCS network using bayesian neural network", *IEEE on Advanced Communication Technology*, Vol. 3, pp. 2185- 2190, February 2009.
[19] B. V. Gokalgandhi, D. Dereyk, A. J. Varia, and S. Samant, "Bandwidth and radiation management of GSM system using neural networks", International Journal of Computer Applications, Vol. 107, pp. 24-26, December 2014.
[20] T. P. Oliveira, J. S. Barbar, and A. S. Soares, "Computer network traffic prediction: A comparison between traditional and deep learning neural networks", International Journal of Big Data Intelligence, Vol. 3, pp. 28-37, January 2016.
[21] J. H. Ni, N. N. Sun, and Z. W. Duan, "Bandwidth prediction research based on BP neural networks in soft switch power dispatching network", IEEE 6th International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII), Vol. 3, pp. 476-478, November 2013.

[22] S. S. Chaudhari, and R. C. Biradar, "Available bandwidth prediction using wavelet neural network in mobile ad-hoc networks", IEEE International Conference on Circuits, Communication, Control and Computing (I4C), pp. 295-299, November 2014.
[23] Y. Jiao, S. Zieli_ski, and F. Rumsey, "Hierarchical bandwidth limitation of surround sound-part II: Optimization of bandwidth allocation strategy", *AES: Journal of the Audio Engineering Society*, Vol. 57, pp. 5-15, March 2009.
[24] H. W. Chu, and D. H. Tsang, "Dynamic bandwidth allocation for real-time VBR video traffic in ATM networks", *IEEE Proceedings of the International Conference on Computer Communications and Networks*, pp. 306-312, September 1997.
[25] J. H. Wang, J. Y. Pan, and Y. C. Cheng, "Session recognition and bandwidth guarantee for encrypted internet voice traffic: Case study of skype", *IEEE Symposium in Computational Intelligence and Data Mining*, pp. 384-389, March 2007.
[26] Hijazi, H. Inoue, A. Matrawy, P.C. Oorschot, and A. Somayaji, "Discovering Packet Structure through Using Li ghtweight Hierarchical Clustering", *IEEE International*