International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 15537-15544

http://www.ripublication.com

Learning and Analyzing User Behavior By Naive Bayesian Classifier and Validating Through Fuzzy Inference Model

Suresh K.S., Nuggu Subba Venkata Gayathri Devi, Routhu Keerthi Praneetha and Kuppam Aswini Durga, Ananthakrishnan S.

Department of Computer Science, School of Computing, SASTRA University

kssuresh@cse.sastra.edu

Abstract

In the age of internet, need of analyzing user behavior is inevitable in various field of businesses. The organizations either cater to the need of users' requirements or to monitor their behavior. In this paper, user's behavior is tracked and Naïve Bayesian Classifier is employed to analyze their behavior depends on their navigation pattern over the usage of applications or their browsing contents. The result is validated by implementing the fuzzy inference method.

Keywords: User behavior, monitoring computer user's activities, Naïve Bayesian Classifier, Fuzzy Inference Method.

Introduction

The need of user behavior analysis can be categorized into two aspects [1]. On the positive side, in the globally growing online business environment in the world of web, includes: Surveying the opportunities for a business, Estimating the market demand for the products, and Understanding the need and grievances of the customers. On the other side, users are monitored to avoid intruder's activities to protect the systems sensitive data and to improve the productivity of the individuals in turn to improve the overall organization's productivity.

In this paper, the individuals are monitored and they are analyzed according to their usage of applications or browsing contents. In an organization environment, even if the users are being monitored, it is unlikely to totally eradicate the unproductive activities. However it has to be observed and evaluated to analyze how much the system users are deviating from their track to be accountable. Time is considered as a primary parameter to measure the productivity whether it is positive or negative depending on the relevancy.

The user behavior analysis involved with two phases as capturing the actions and verification of actions. The latter phase can be done either in a static manner or in a dynamic manner. In the static approach the users' actions are tracked and recorded in a database then the actions are scientifically quantified to measure the user behavior. In the case of dynamic approach, the users' behavior is analyzed dynamically where actions are to be evaluated on the fly and respond for the actions of the users, whether it is case of intruder detection or on line business purpose.

The Naïve Bayes classifier [2] is used to estimate the grade of the users based on their usage of applications. To validate the work, Fuzzy Inference Methodology is implemented through MATLAB for the samples taken and evaluated to find the behavior. It is evident that the model is working well and proved by mapping the result with the derived grade from Naïve Bayes classifier.

Related Works

The foundation for the discussion started by Quinlan J R in late 70s in his "Discovering Rules by Induction From Large Collections of Examples". And his contribution continued in this field of study [3],[4],[5],[6]. The next significant role was played by J.C. Schlimmer and D.H. Fisher for the field of capturing history and cost effective learning method of behaviour in [7] discussed in the paper about the dynamically updating activities considered as the input data and was processed to analyze and derive the behaviour of the user.Macedo et al. [8], proposed a system to define the behavior of a list of existing users from the available activities as history. Pepyne et al. [9] recommended using queuing theory and logistic regression modeling methods those who have the similar way of usage. The approach with intelligent agents [10] is used for the preferences of the users. The users' profile may be evolved dynamically so that to find the intruders which requires automatic detection with effective user profiles [11].

The popular methodologies such as Decision Tree Learning (DTL), Bayesian Classifier (BC), Artificial Neural Networks (ANN), Learning Vector Quantization Algorithm (LVQ), Dynamic Vector Quantization Algorithm (DVQ) and Support Vector Machine (SVM) are employed in the field learning, predicting and evaluating the user's behaviors depending on the application requirement.

Decision Tree Learning

This concept is used as incremental decision tree learning to take decision on dynamically updating history.

Naive Bayesian Classifier

This is a probabilistic type of classifier and the prediction has to be done considering all the features of class simultaneously.

Artificial neural networks (ANN)

Kasabov [12] devised a model called Evolving Fuzzy Neural Network . According to this, without accessing previous data, the new class will be created.

LVQ and DVQ

These two methods are categorized under prototype-based supervised algorithm. Based on this principle, the prototypes are constructed to categorize the data set. The occurrence of new instance is attached with the most similar prototype class by comparing the instance with the existing prototype. The comparison is done be distance measuring method.[13,14]

SVM

The principle behind this concept is, categorizing the data into two with old and new. Then the new data is incrementally trained to predict the behavior. Xiao et al.[15] suggested a technique through the adjustable parameters by adding distribution history exploiting the SV set properties.

Justification For Using Naïve Bayes Classifier

The different classifiers [16] and classification methodologies [17] are experimented for effective user profiling. The significance of Naïve Bayes classifier performs so efficiently in a supervised learning environment using trained set. By considering and estimating the many independent attributes, the prediction can be done very competently to determine the class.

In this specific case, the converted percentile forms of the time duration of the application utilized by the users are considered as different attributes. This is mapped into three different range or features and then applying Bayesian principle to obtain the grade for the individual users. When the number of applications and the number of users increases, manual calculation will become so complicated. But the Bayes classifier can be implemented to find the solution in a complex situation.

Phases of the Implementation

As shown in the figure 1, the implementation consists of four phases such as

1. Capturing Users activities,
2. Conversion and Classification
3. Graphical Representation and
4. Validation

The classifier implementation is to learn the behavior of the user and evaluate the activities related to time. The first three phases are implementation and last phase for validation using fuzzy inference method.

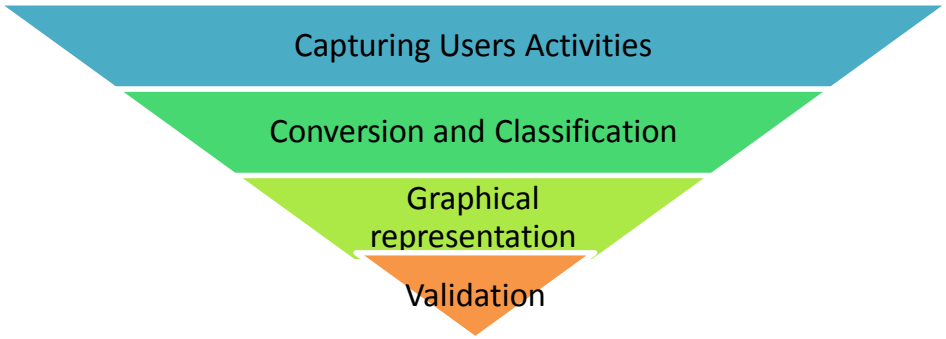


Figure 1: Phases of implementation and validation

Implementation of Phases

The first three phases are implemented through java and My SQL. For validation the matlab has been used to implement the fuzzy inference rule.

Phase-1: Capturing Users Activities Phase

The Users credentials are kept in a database to verify the user's login and password. In addition to these static credential fields, there is a room to maintain the dynamic details with the field names as the applications provided in the virtual desktop to save the time span spend by the individuals on the given applications for testing. Arbitrarily the different applications provided are notepad, netbeans, google, youtube, facebook etc.. the applications to be added may be decided by the specific organization's requirement.

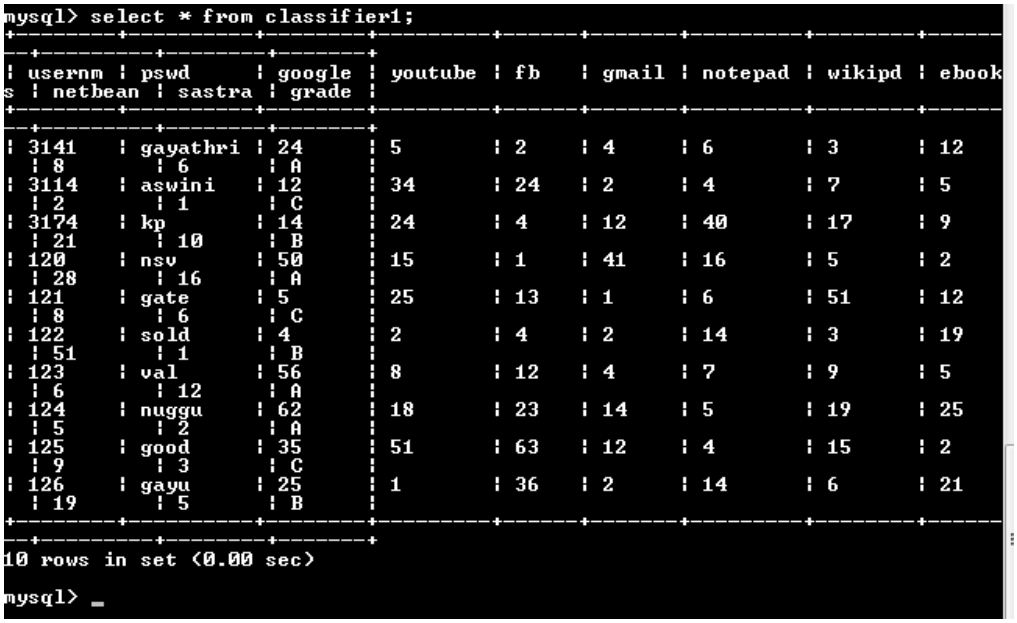
After verifying the entered user name and password in the interface provided to the individual users, virtual desktop will be available to select any application. The specific application usage time are written into the database for the users by calculating time span between opening and closing of the applications, thanks to java classes with useful methods(exec( ) - Runtime class, wait For( ) - Process class, get Time( ) - Date class methods) to capture the information with time. So each record contains the individual's user name and password and their corresponding time used for each application in minutes which is going to be used for the classification phase.

Phase-2: Conversion and Classification Phase

Conversion Phase

Before applying classification methodology, the percentile conversion is done with the data stored in the field of applications. The duration of each application used by the user is taken and they are summed up and converted into 100 percentile form and modified as shown in the figure 2.

$$\text{Percentile value} = \frac{\text{Time span of each application} * 100}{\text{Total duration}}$$



```
mysql> select * from classifier1;
+--------+----------+--------+---------+------+-------+---------+--------+------+
| usernm | pswd     | google | youtube | fb   | gmail | notepad | wikipd | ebook
s | netbean | sastra | grade |
+--------+----------+--------+---------+------+-------+---------+--------+------+
| 3141   | gayathri | 24     | 5       | 2    | 4     | 6       | 3      | 12
  | 8       | 6      | A     |
| 3114   | aswini   | 12     | 34      | 24   | 2     | 4       | 7      | 5
  | 2       | 1      | C     |
| 3174   | kp       | 14     | 24      | 4    | 12    | 40      | 17     | 9
  | 21      | 10     | B     |
| 120    | nsv      | 50     | 15      | 1    | 41    | 16      | 5      | 2
  | 28      | 16     | A     |
| 121    | gate     | 5      | 25      | 13   | 1     | 6       | 51     | 12
  | 8       | 6      | C     |
| 122    | sold     | 4      | 2       | 4    | 2     | 14      | 3      | 19
  | 51      | 1      | B     |
| 123    | val      | 56     | 8       | 12   | 4     | 7       | 9      | 5
  | 6       | 12     | A     |
| 124    | nuggu    | 62     | 18      | 23   | 14    | 5       | 19     | 25
  | 5       | 2      | A     |
| 125    | good     | 35     | 51      | 63   | 12    | 4       | 15     | 2
  | 9       | 3      | C     |
| 126    | gayu     | 25     | 1       | 36   | 2     | 14      | 6      | 21
  | 19      | 5      | B     |
+--------+----------+--------+---------+------+-------+---------+--------+------+
10 rows in set (0.00 sec)

mysql>
```

Figure 2: Sample table for usage of applications.

To apply the Naïve Bayesian, the calculated Percentile values are categorized into three levels named as low (0-30), medium (30-60) and high (60-100). The training set is generated according to the rules framed depending on the organizational expectations and stored in database initially as shown in the figure 3.

Google	Youtube	Facebook	Grade
High	Low	Medium	A
High	Medium	Low	A
Medium	Medium	Low	B
Medium	Medium	Medium	B
Low	High	High	C
Low	High	Medium	C

Figure 3: Training set with Grade to apply Naïve Bayesian method

Implementation of Bayesian classifier

To calculate the posterior probability value for each grade, the following expressions are used.

$$P(X) = \frac{\text{Probability of Grade X}}{\text{Total Number of Data Sets}}$$

where X=Grade A, Grade B, Grade C

P (application=Percentile value | X) = Number of occurrences of the application and its value for the given Grade X.

where application = google, youtube, fb etc.,

Percentile value = low, medium, high

For Example,

$P(\text{youtube} = \text{low}|A) = \frac{1}{2}$

Finally the Posterior probability of each grade is calculated with the formula

$$Pos(X) = P(X)\ x\ P\left(\frac{google}{X}\right)\ x\ P\left(\frac{facebook}{X}\right)\ x\ P\left(\frac{youtube}{X}\right)$$

And the highest posterior probability is the resultant grade of the user and it is updated into the database to its respective user by Prepared Statement.

The Administrator will have the privilege to view each user's behavior by giving their respective User ID. The grade of the user estimated by the system will exemplify the productivity of the user. These data are extracted from the database.

Phase -3: Graphical Representation Phase

The pie chart is generated to visualize the relative time span used by the users on the specific applications to understand the productivity of the users by taking data from database and the values are summed and start angle and arc angle are calculated by

$$\text{startAngle} = \left(\frac{\text{curValue} * 360}{\text{total}}\right)$$

$$\text{arcAngle} = \left(\frac{\text{slices[i].value} * 360}{\text{total}}\right)$$

These values are used in fillArc() method to draw the arc of each value.

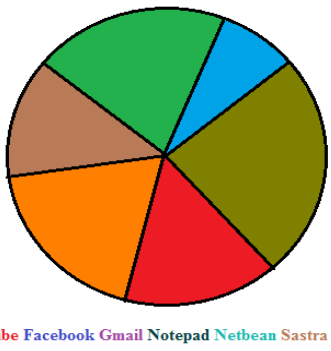


Figure 4: Relative time spent on different applications

Phase-4: Validation Phase

In matlab, Fuzzy Inference Method with appropriate Membership functions is used to validate the obtained solution by using Naïve Bayesian Classifier. The input and

output variables are defined and each variable is classified into low, medium and high. The rules are defined according to the input and output variables to check the particular grade and it can be viewed by View Rules. In the matlab, the value is produced for individual users are compared and mapped to the corresponding grade. It is verified that the value obtained by fuzzy method and the grade calculated by Naïve Bayesian classifier are same.

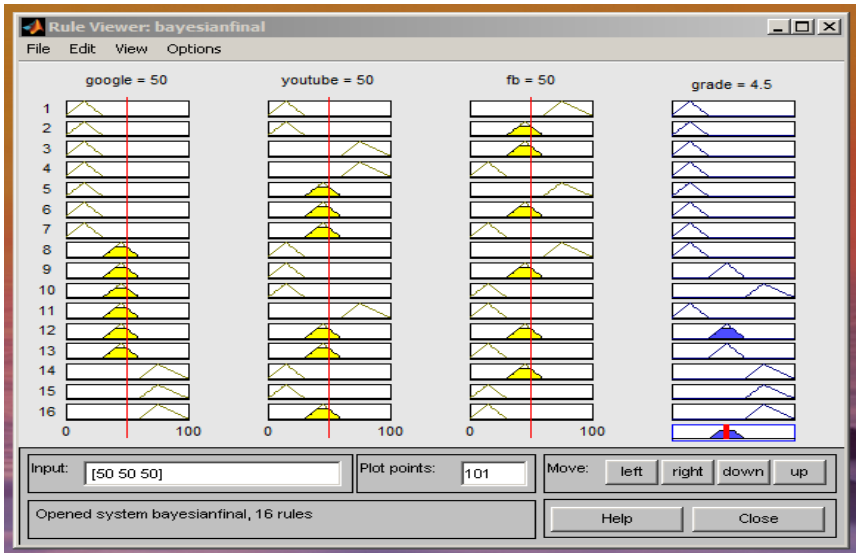


Figure 5: Output of Fuzzy Inference Method for validation

Conclusion

The user's activities are monitored and evaluated to increase the productivity. But the great advantage of this system is, instead of absolutely restricting the accessing or browsing unproductive applications or sites, it allows, monitors and evaluates the productivity and categorizes the users with grade. Hence the actions to be taken can be done in an appropriate and scientific manner which may not disturb the total freedom and privacy of the user in an organizational environment.

References

[1]. Iglesias JA, Angelov P, Ledezma A, de Miguel AS, 2012 Creating evolving user behavior profiles automatically. IEEE Trans Knowl Data Eng 24(5):854–867

[2]. Shin YE, Choi WH, Shin TM., 2014,Physical activity recognition based on rotated acceleration data using quaternion in sedentary behavior : A preliminary study Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE

[3]. Quinlan,J.R."Discovering Rules by Induction From Large Collections of Examples", 1979

[4]. Quinlan, J. R.Inferno: A Cautious Approach To Uncertain Inference. 1983

[5]. Quinlan, J. R. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106, 1986

[6]. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[7]. J.C. Schlimmer and D.H. Fisher, "A Case Study of Incremental Concept Induction," Proc. Fifth Nat'l Conf. Artificial Intelligence (AAAI), pp. 496 501, 1986.

[8]. A. Alaniz Macedo, K.N. Truong, J.A. Camacho Guerrero, and M. Graca Pimentel, "Automatically Sharing Web Experiences through a Hyperdocument Recommender System," Proc. ACM Conf. Hypertext and Hypermedia (HYPERTEXT '03), pp. 48 56, 2003.

[9]. D.L. Pepyne, J. Hu, and W. Gong, "User Profiling for Computer Security," Proc. Am. Control Conf., pp. 982 987, 2004

[10]. D. Godoy and A. Amandi, "User Profiling in Personal Information Agents: A Survey", Knowledge Eng. Rev., vol. 20, no. 4, pp. 329 361,2005.

[11]. Jose Antonio Iglesias,Plamen Angelov, Agapito Ledezma, and Araceli Sanchis, "Creating Evolving User Behavior Profiles Automatically", KNOWLEDGE AND DATA ENGINEERING VOL.24 NO.5 YEAR 2012.

[12]. Kasabov, N., "Evolving fuzzy neural networks for supervised/unsupervised online knowledge-based learning", Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on (Volume:31 , Issue: 6)

[13]. T. Kohonen, J. Kangas, J. Laaksonen, and K. Torkkola, "Lvq pak: A Program Package for the Correct Application of Learning Vector Quantization Algorithms," Proc. IEEE Int'l Conf. Neural Networks, pp. 725 730, 1992.

[14]. F. Poirier and A. Ferrieux, "Dvq: Dynamic Vector Quantization An Incremental Lvq," Proc. Int'l Conf. Artificial Neural Networks, pp. 1333 1336, 1991.

[15]. R. Xiao, J. Wang, and F. Zhang, "An Approach to Incremental SVM Learning algorithm," Proc. IEEE Int'l Conf. Tools with Artificial Intelligence, pp. 268 278, 2000.

[16]. A. Cufoglu, M. Lohi, and K. Madani, "A Comparative Study of Selected Classifiers with Classification Accuracy in User Profiling", Proc. WRI World Congress on Computer Science and Information Eng. (CSIE), pp. 708 712, 2009.

[17]. T. Cover and P. Hart, "Nearest Neighbor Pattern Classification", IEEE Trans. Information Theory, vol. 13, no. 1, pp. 21 27, Jan. 1967.