INTERPRETABLE RECOMMENDATIONS FOR NORMAL BEHAVIOR AND DETECT FRAUD IN GRAPHS WITH TIME

[1] Jeevanandhini.P
[1] Research Scholar ,
[2] A. Nirmala, MCA, MPhil(Ph.D)
[2] Assistant Professor,
[1&2] Dr. NGP arts and science college , TamilNadu , India.

Abstract

In this paper we examine a data mining approach for taking in probabilistic client behavior models from the database use logs. We propose a methodology for making an interpretation of database follows into portrayal appropriate for applying data mining methods. To stay away from this issue we propose novel method dependent on mix of choice tree classification algorithm and empirical time-subordinate component outline, by potential functions theory. The execution of the proposed method was experimentally evaluated on real-world data. The correlation with existing cutting edge data mining methods has affirmed remarkable execution of our method in prescient client behavior modeling and has exhibited competitive outcomes in anomaly detection.

Keywords:

1. Introduction

As "Big Data" has turned out to be inescapable, associations in each industry are putting away the greatest number of activities and communications as they can. For both scholarly community and industry these huge databases have huge undiscovered potential, with expanding interest in attempting to find valuable examples. These databases would all be able to be seen as vast hyper graphs among elements of various sorts with numerous attributes contextualizing their connections. Through modeling these communications, software engineering can give insights past what we can see and see individually and push numerous fields forward. With more data being put away than any other time in recent memory, we currently have the chance to move past just foreseeing whether two substances have associated yet in addition understanding the setting of those communications dependent on the numerous important attributes accessible. This fast development in the kinds of co-operations and contextual data being put away displays another boondocks for graph modeling, empowering new applications and exhibiting new challenges in data mining and scalable machine learning. To give a few precedents: online clients associate with each other in social networks as well as with their general surroundings supporting government officials, watching motion pictures, purchasing clothing, searching for eateries and notwithstanding discovering specialists. These

associations regularly incorporate insightful contextual data as attributes, such as the time or area of the cooperation and appraisals or surveys about the collaboration. There are comparative hyper graphs in healthcare associations among patients, specialists, illnesses and medications, and also in educational communications among understudies, teachers, subjects and educational resources. In each of these fields, researchers have possessed the capacity to extract helpful knowledge just from the graph structure, e.g., anticipating what films a man will like and discovering networks of individuals with comparative interests. However, to give more insightful comprehension, we have to model the associations as well as the setting of the connections, finding the important attributes within a graph and saturating our models of those attributes with our instinct. Making utilization of all of these data successfully shows many energizing research challenges, running from designing models that catch the pertinent examples in huge, attributed hyper graphs to frameworks for real-world challenges, we can create holistic arrangements that augment real-world effect. To do this, we connect insights from applications, modeling and scalable machine learning frameworks.

Top recommender frameworks have utilized thousands of components for every thing and per client, similar to the case in the triumphant entries in the Netflix prize. Late best in class methods have depended on adapting much bigger, more perplexing factorization models, frequently taking nontrivial blends of different submodels. Such perplexing models are progressively hard to interpret, utilize a lot of memory, and are frequently troublesome coordinate into bigger frameworks. Co-bunching then again accept there exists some "right" dividing of motion pictures (and clients). For example, a client might be a piece of a gathering that likes all comedies yet does not like sentimental motion pictures. Correspondingly, all rom-coms might be part out into a different bunch. In the event that a gathering of clients likes new films however not old ones, each kind might be further parceled by decade. This quickly prompts a combinatorial blast. By taking a straight mix of co-clusterings we profit by both points of view: by modeling the discrete idea of attributes we can keep away from the expense of high-dimensional factorization models, and by including the inclinations for various attributes we can stay away from the vast models important to cover all blends of attributes. Rather, through backfitting, we make an all the more amazing, hierarchical portrayal.

2. Literature Survey

1. **Liang Zhao, Zhikui Chen, Yueming Hu, Geyong Min and Zhaohua Jiang**(2014) proposed framework for productive analysis of high-dimensional financial huge data dependent on imaginative disseminated include determination. In particular, the framework joins the techniques for monetary element determination and econometric model development to uncover the concealed examples for financial

improvement. The usefulness lays on three columns: (I) novel data pre-handling techniques to get ready high-quality monetary data, (ii) a creative dispersed element distinguishing proof answer for find vital and delegate financial markers from multidimensional data sets, and (iii) new econometric models to catch the concealed examples for monetary improvement. To start with, fundamental components are connected to depict the instrument of financial development. The monetary development can be advanced by expanding utilization and venture, and in addition influencing related unequivocal variables. When moving toward monetary analysis, the contributing variables are chosen to recognize the relations among them and financial improvement. Second, from the point of view of cost sparing, urbanization can bring more workforces into city, which lessens the financial expenses and lifts offices sharing to chop down exchange costs. 2. **Mohammad Shorfuzzaman(**2017) proposed the effect of big data examination on knowledge the board and proposes a cloud-based theoretical framework that can break down big data progressively to encourage upgraded basic leadership planned for upper hand big data because of its different properties like high volume, assortment, and speed can never again be successfully put away and investigated with conventional data the executives techniques . New technologies and designs are required to store and break down this data and thus produce indispensable ongoing information for basic leadership in associations. This has opened the entryway for the scientists to center around big data examination which are probably going to assume an extreme job in the accomplishment of associations. The test is to gather, store, and break down the undertaking big data at the correct speed from sources, for example, deals, inventory network, research, and client relations to manufacture the knowledge base for compelling basic leadership of the associations. Ongoing investigations additionally uncovered the way that the usage of big data has enlarged prominently in basic leadership and both open and private associations are receiving rewards from this rising innovation. There is no endless supply of big data accessible in the writing. 3. **Dr. Venkatesh Naganathan**(2018) proposed Big data its difficulties and its future degree where it is driving as well and Big Data Analytics strategies utilized by various associations that encourages their business to settle on a solid speculation choices. Data and examination are at the core of the advanced upheaval. They are a basic over all enterprises. To endure and flourish in the advanced period, right now is an ideal opportunity to drive data and examination into the center of your business and scale outward to each worker, client, provider, and accomplice. Scaling the estimation of data and investigation requires a culture of data enablement that reaches out all through each aspect of your association. A culture where data and investigation advise and drive business objectives, operational efficiencies, and development from Gartner Data and Analytics Summit. 4. **RakeshRanjan Kumar&BinitaKumari**(2015) proposed Big Data mining, issues identified with mining and the new chances. DATA MINING PARAMETERS The most usually utilized techniques in the data mining are: Association - Looking for examples where one occasion is associated with another occasion. Counterfeit neural networks - Non-straight prescient models that learn through preparing and take after natural neural networks in structure Classification - is a systematic procedure for acquiring imperative

and important information about data, and metadata – data about data. Clustering - the way toward recognizing data sets that are like each other to comprehend the distinctions and in addition the similitudes inside the data. Choice trees: Tree-molded structures that speak to sets of choices. 5.**Vivekananth.P , Leo John Baptist.A**(2015)proposed distinctive data investigation techniques, for example, text examination, audioanalytics, videoanalytics, online networking investigation and prescient examination The resultant impact of having such an enormous measure of data will be data investigation. Data examination is the way toward organizing big data. Inside big data, there are distinctive examples and relationships that make it workable for data examination to improve ascertained portrayal of the data. This makes data examination a standout amongst the most critical parts of information innovation. There are diverse techniques that are at present being used. As a rule, the techniques can be summed up to: 1) Association Rule Learning 2) Classification tree analysis 3) Genetic calculations 4) Machine learning 5) Regression analysis 6) Sentimental Analysis 7) Social network analysis.

3. Methodology

3.1 Synchronized Behavior Detection

It has been established by past works that numerous data mining approaches on graphs profit by abusing the highlights from the nodes' behavior designs, including (an) out degree and in-degree, (b) HITS score (hubness and authoritativeness), (c) betweenness centrality, (e) node in-weight and out-weight, if the graph is weighted, (f) the score of the node in the I-th left or right solitary vector, and some more. We signify k-dimensional component vector of node u by p(u) 2 Rk. We extract the element vector from graph structure that somewhat mirrors the node's behavior design. The highlights could be any from the abovementioned and the vector could have any dimensionality. In this chapter, we choose the degree values and HITS score. We signify a set of u's source nodes by I(u) and a set of u's target nodes by O(u). The in-degree di(u) of node u is the number of its sources, i.e. the span of I(u). The out-degree do(u) of node u is the number of its targets, i.e. the extent of O(u). Also we signify by hub(u) the hubness of node u and by aut(u) the authoritativeness of u, as per Kleinberg's popular work. We choose these highlights for two reasons: they are quick to figure, and in addition simple to plot. As our investigations show, they work well in stick pointing suspicious nodes. Note that if the side data is accessible, it could be viewed as additional highlights that would be effortlessly fused, and hopefully, the execution could be better.

We give proof that CatchSync is compelling at both the classic issue of marking suspicious behavior, and also surfacing new examples of unusual gathering behavior:

Detection effectiveness: We demonstrate CatchSync’s ability to accurately label suspicious behavior and remove anomalies through three techniques.

1. Injected attacks: We begin by testing the accuracy, precision, and recall on synthetic graphs with injected group attacks. We compare our algorithm against state-of-the-art methods and show that CatchSync performs the best.

2. Labelling task: We also test our accuracy, precision, and recall on two realdatasets, where we use the labeled data from random sampling of TWITTERSGand WEIBOSG as ground truth of suspicious and normal nodes.

3. Restore normal patterns: For all of these cases we show that removing the suspicious nodes restores the power law properties of the graph's edge degree, which when distorted is a common sign of spam, and remove anomalous, patterns in the feature spaces (OutF-plots and InF-plots).

CatchSync properties: We test a number of properties of CatchSync, including the robustness with respect to, the speed and the scalability.

Discovery: We demonstrate the effectiveness of CatchSync as a discovery tool. We discuss a number of the unusual accounts caught and patterns detected in the TWITTERSG and WEIBOSG datasets.

4. Proposed Work

4.1 Detect Fraud in Graphs With Time

Serial Algorithm

Since we currently might want to run the algorithm for numerous bunches in parallel, we should present some additional documentation. We characterize s to be the number of bunches being run at the same time. Each group has an inside c(k) 2 RM and a set of right now chosen sections P0 k _ P (each characterized as previously). We characterize C to be the set of all c(k) and P to be the set of all P0 k , both for k = 1 : s. Like the serial algorithm, the MapReduceCopyCatch algorithm works by refreshing c and P0 iteratively. The center of the algorithm can be found in Algorithm 3, where we take note of that we run one MapReduce work for each cycle, each time refreshing C and P. As in the serial algorithm, we can keep iteratively refreshing C and P until the point when no changes are made. By and by, we will only run a settled number of emphasess.

MapReduceIt is worth taking a minute to take note of the data stream in a MapReduce work before depicting the details of our algorithm. In the Map step, our information is part among numerous mappers. Each mapper gets a couple of data of the frame hKEYmap; VALUEi and can yield at least zero consequences of the shape hKEYreduce; VALUEi. In the reducer venture, for each exceptional KEYreduce a reducer is framed which takes as an info hKEYreduce; VALUESi, where VALUES is a set of the VALUE yields from the mapper step which compare to that reducer's specific KEYreduce. The reducer would then be able to yield data to disk. Beside this data stream, we make utilization of Hadoop's Distributed Cache, which lets us store data as global read-just data.

USERMAPPER discovers which clients are as of now in which groups dependent on C and P and maps those clients to a reducer dependent on which bunch it is within. All the more specifically, the Map

step takes as information L and I, where each (Li;_; Ii;_) for all I 2 U is contribution to a mapper. Each mapper checks the Li;_ over all s groups to check whether it falls within that bunch following the definition given in our advancement objective (4.10). In the event that it does, it discharges a yield where the key is the bunch ID k, and the value is the column (client) data Li;_ and Ii;_. Each mapper keeps running in O(sm) time, and since this is being kept running over all data the whole advance takes O(smN) not taking into record parallelization.

AdjustCluster-Reducer takes in all of the clients as of now in a given bunch k and updates c(k) and P0 k. Each reducer takes as an info hk; U0i, where U0 here contains sets (Li;_; Ii;_) for all clients in the bunch (as was yield by the mappers). As shown in Procedure 5, we should be mindful so as to just utilize values from each client in the measurements for which it falls within the group. However, past this, the refresh generally works reasonably basically. The middle c is refreshed by simply taking a normal of the focuses in the group, like a mean-shift algorithm with a level kernel. The chose sections are chosen dependent on which segments cover the most clients from the past bunch parameters, and afterward by which segments have the least change among these clients. By utilizing the past communities for this refresh, we can do the calculation online, disregarding each client just once. Since we accept each group is O(n) in size, each reducer takes O(nM) time, and the diminish venture as a whole takes O(snM) when not thinking about parallelization. The reducers yield the refreshed C and P to be set in the Distributed Cache in resulting cycles.

Seeds and Initial Iterations Figuring out where to begin the bunches can be exceptionally troublesome. To stay away from any inclination, we test seeds arbitrarily from the rundown of all edges in the graph, utilizing the edge's Page and Like time to initialize both P0 and c. (While we could utilize suspicious clients from one of Facebook's numerous other security mechanisms, this would present earlier presumptions about attackers that are superfluous and could make it less demanding for an enemy to hide.) subsequently, in the initial cycles clients just need Liked a solitary Page at around a similar time, which isn't extraordinary. There are frequently such a significant number of clients found in these initial emphasess that we test a small level of them at arbitrary to keep the algorithm productive. Indeed, even with the testing, this method lets us quickly discover loads of clients that Liked one Page around a similar time, and afterward observe what else they have in like manner. This testing is performed in the initial two cycles until jP0j = m.

4.2 Interpretable Recommendations for Normal Behavior

The mathematical challenge that propelled this work is, how would we be able to make a concise model of client behavior that still gives high quality expectations? Designing a compact model is troublesome in light of the fact that it requires making suspicions and limitations while not diminishing the model's exactness. Factorization models are entirely adaptable. With the end goal to encode a rank-k framework by a factorization, we require k numbers per line (and segment) separately. Rather, think about a stencil - a small k _ k layout of a framework and its mapping to the line and segment vectors separately. We

just need log2 k bits per line (and section) in addition to O(k2) drifting point numbers paying little mind to the measure of the network. Taking a direct mix of stencils we can model very mind boggling grids. This is best comprehended by the model beneath: expect that we have two straightforward stencils containing 3 _ 2 and 2 _ 3 co-clusterings. Their direct mix yields a rather nontrivial 9 _ 8 lattice of rank 5. Alternatively, classic co-bunching would require a (3 _ 2) _ (2 _ 3) parceling to match this structure. When we have S stencils of size k _ k, this would require a solitary co-grouping of size kS _ kS.

Algorithm

We consider a simple iterative procedure to learn stencils to approximate our data R.Algorithm 1gives the high level algorithm of learning each stencil one at a time. Inlearning each stencil we use the CLUSTER algorithm, similar to k-means, as given inAlgorithm 2.

Algorithm 1

Require: matrix $\mathbf{R}$, indicator matrix $\mathbf{I}$, clusters k_n, k_m, max stencils S

$\hat{\mathbf{R}} \leftarrow \mathbf{R}$
for $\ell = 1$ **to** S **do**
  $(\mathbf{c}^{(\ell)}, \mathbf{V}^{(\text{row})}, \mathbf{L}) \leftarrow \textsc{Cluster}(\hat{\mathbf{R}}, k_n, \mathbf{I})$ {Rows}
  $(\mathbf{d}^{(\ell)}, \cdot, \cdot) \leftarrow \textsc{Cluster}([\mathbf{V}^{(\text{row})}]^\top, k_m, \mathbf{L}^\top)$ {Columns}
  for all $a, b \in \{1, \ldots k_n\} \times \{1, \ldots k_m\}$ **do**
    $\mathbf{T}^{(\ell)}_{a,b} \leftarrow \text{mean}\left\{\hat{\mathbf{R}}_{u,m} | \mathbf{c}^{(\ell)}_u = a \text{ and } \mathbf{d}^{(\ell)}_m = b\right\}$
  end for
  $\hat{\mathbf{R}} \leftarrow \hat{\mathbf{R}} - \mathcal{S}(\mathbf{T}^{(\ell)}, \mathbf{c}^{(\ell)}, \mathbf{d}^{(\ell)})$ {Backfit on residuals}
end for
return $\{\mathbf{T}^{(\ell)}, \mathbf{c}^{(\ell)}, \mathbf{d}^{(\ell)}\}_{\ell=1}^{S}$

Row clustering We first perform k-means clustering of the rows. That is, we aim to findan approximation of R that replaces all rows by a small subset thereof. Algorithm 2isessentially a generalization of k-means clustering. By calling the algorithm withM = R,k = kn, and W = 1N_M, we find that the algorithm simplifies significantly to classick-means.

Algorithm 2

Once we have run the clustering algorithm on the rows, we nowcluster the columns of previously learned row clusters, V(row). In this case, we need toweight each row of V(row) (corresponding to a row cluster) by the number of rows itrepresents (the number of rows in that cluster). As a result, rather than choose a columncluster assignment by the Euclidean distance, we use the Mahalanobis distance. Still,Algorithm 2is a generalization of this concept.

Require: matrix $\mathbf{M} \in \mathbb{R}^{N_1 \times N_2}$, weights $\mathbf{W} \in \mathbb{R}^{N_1 \times N_2}$, number of clusters k

Draw k rows from $\mathbf{M}$ at random without replacement and copy them to $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\} \in \mathbb{R}^{k \times N_2}$.
while not converged **do**
  $\mathbf{L} \leftarrow 0 \in \mathbb{R}^{k \times N_2}$ and $\mathbf{Y} \leftarrow 0 \in \mathbb{R}^{k \times N_2}$
  for $i = 1$ **to** N_1 **do**
    $\mathbf{c}_i \leftarrow \text{argmin}_c \sum_j \mathbf{W}_{i,j}(\mathbf{M}_{i,j} - \mathbf{V}_{c,j})^2$
    $\mathbf{Y}_{\mathbf{c}_i,:} \leftarrow \mathbf{Y}_{\mathbf{c}_i,:} + \mathbf{M}_{i,:}$ {Increment statistics}
    $\mathbf{L}_{\mathbf{c}_i,:} \leftarrow \mathbf{L}_{\mathbf{c}_i,:} + \mathbf{I}_{i,:}$ {Increment counts}
  end for
  for $c = 1$ **to** k **do**
    $\mathbf{V}_{c,:} \leftarrow \mathbf{Y}_{c,:}/\mathbf{L}_{c,:}$ {New cluster center}
  end for
end while
return cluster assignments $\mathbf{c}$, clusters $\mathbf{V}$, counts $\mathbf{L}$

5. Experimental Results

5.1 Detect Fraud in Graphs With Time

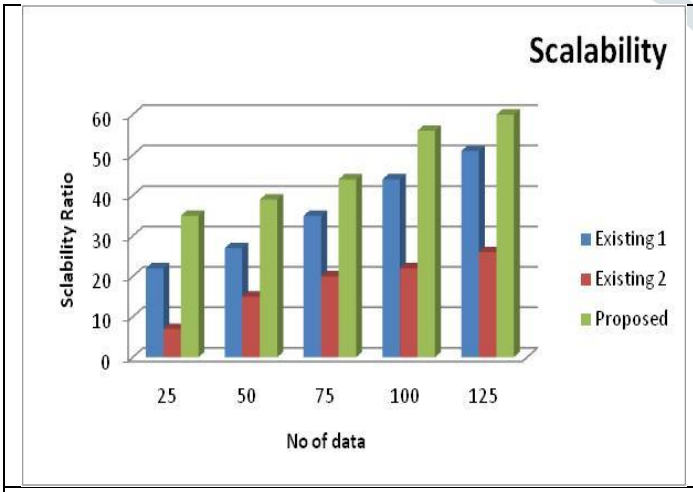

Figure 1: Comparison chart of Scalability

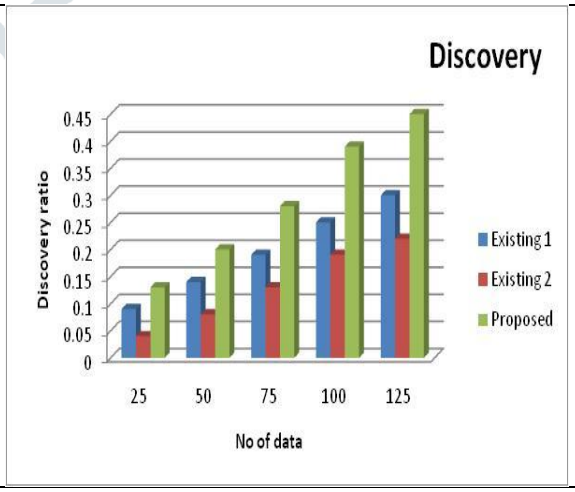

Figure2: Comparison chart of Discovery

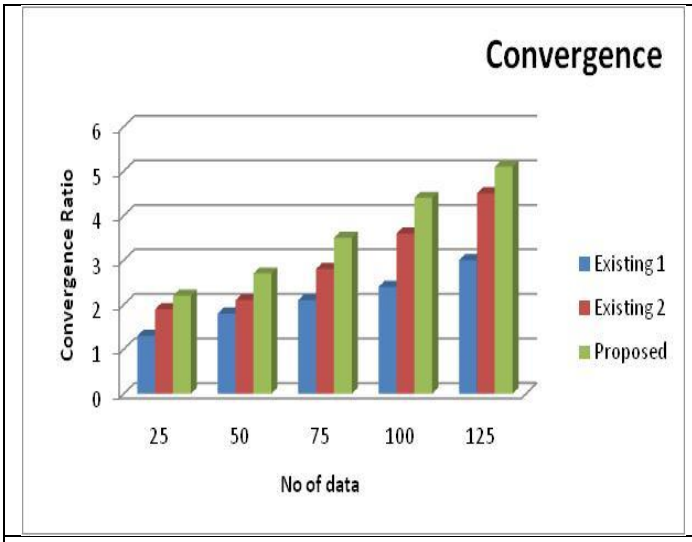

Figure 3: Comparison chart of Convergence

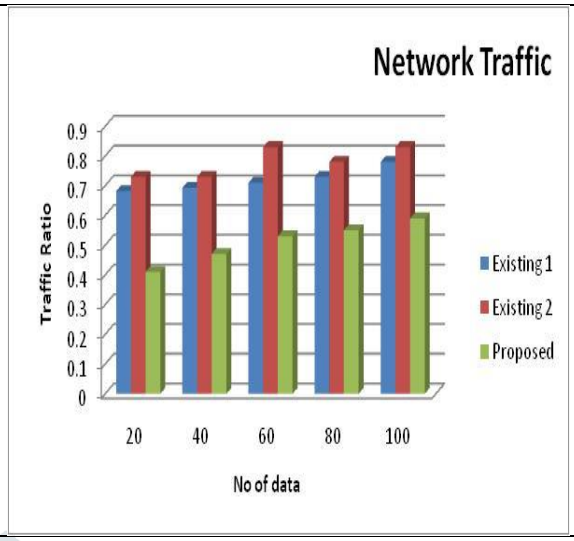

Figure 4: Comparison chart of Network Traffic

The comparison chart of scalability explains the values of existing and proposed method. Scalability ratio in x axis and no of data in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 values are 21 to 51 existing 2 values are 7 to 26 and proposed method values are 35 to 60. The comparison chart of discovery explains the values of existing and proposed method. Discovery ratio in x axis and no of data in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 values are 0.09 to 0.3 existing 2 values are 0.04 to 0.22 and proposed method values are 0.13 to 0.45. The comparison chart of discovery explains the values of existing and proposed method. Convergence ratio in x axis and no of data in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value are 1.3 to 3 existing 2 values are 1.9 to 4.5 and proposed method values are 2.2 to 5.1. The comparison chart of Network Traffic explains the values of existing and proposed method. Network Traffic ratio in x axis and no of data in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value are 0.682 to 0.78 existing 2 values are 0.73 to 0.83 and proposed method values are 0.41 to 0.59.

5.2 Interpretable Recommendations for Normal Behavior

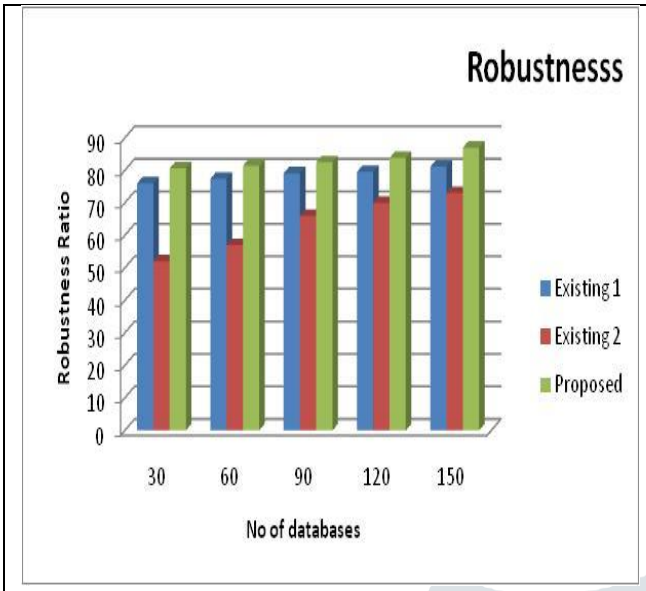

Figure 5: Comparison chart of Robustness

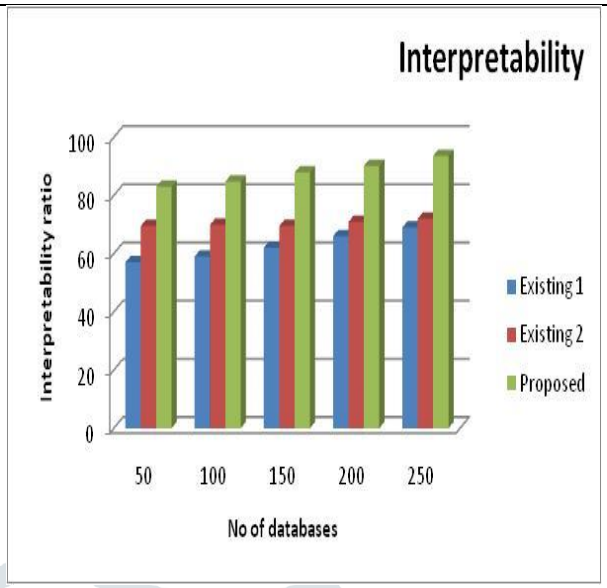

Figure: Comparison chart of Interpretability

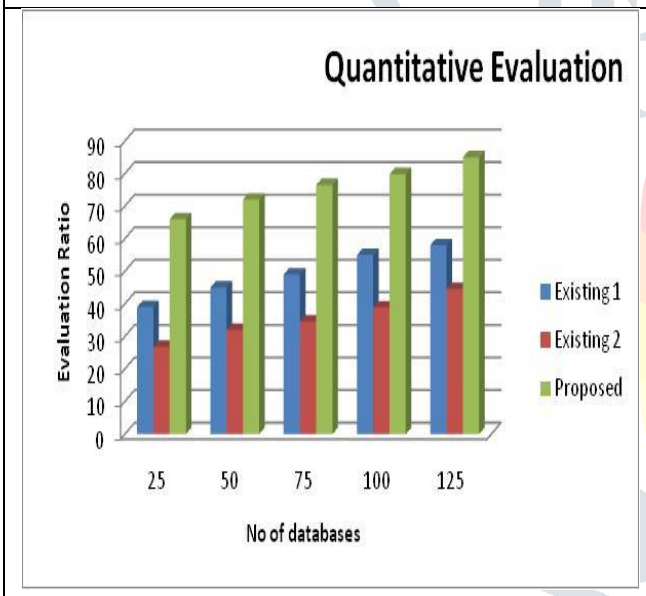

Figure: Comparison chart of Quantitative evaluation

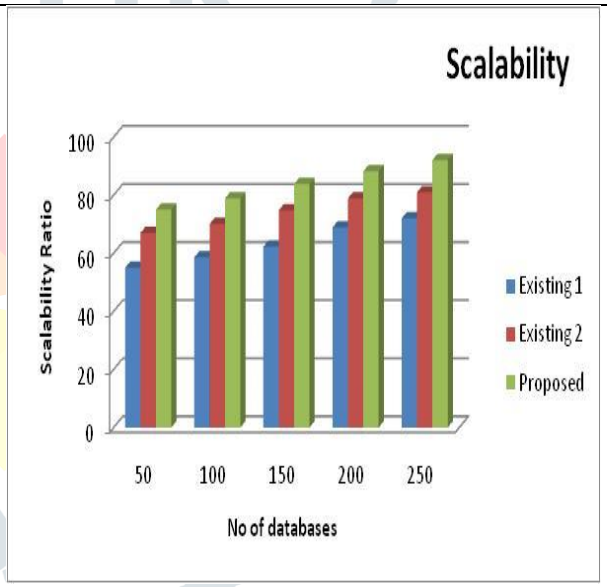

Figure: Comparison chart of Scalability

The comparison chart of robustness explains the values of existing and proposed method. Robustness ratio in x axis and no of data in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 values are 76 to 81 existing 2 values are 52 to 73 and proposed method values are 80.6 to 86.98. The comparison chart of interpretability explains the values of existing and proposed method. Interpretability ratio in x axis and no of databases in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value are 57 to 69 existing 2 values are 69.5 to 72 and proposed method values are 83 to 93.6. The comparison chart of Quantitative Evaluation explains the values of existing and proposed method.

Evaluation ratio in x axis and no of databases in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value are 39 to 58 existing 2 values are 26.77 to 44.56 and proposed method values are 66 to 85. The comparison chart of Scalability explains the values of existing and proposed method. Scalability ratio in x axis and no of databases in y axis. While compare the existing method and proposed method the proposed method gives the better results. Existing 1 value are 55 to 72 existing 2 values are 67 to 81 and proposed method values are 75 to 92.06.

Conclusion

In this paper, we have taken a graph-based point of view to comprehension, modeling, and foreseeing client behavior. Taken together, the work has pushed forward the boondocks of client behavior modeling in various ways, with both scholastic commitments and more extensive effect. We give an outline of these focuses underneath.. Novel method for mining probabilistic client behavior models has been detailed. Unlike other existing data mining methods it consolidates time highlight in the client model. The empirical element outline, by potential functions theory, has been proposed for that. Consolidating this component delineate choice tree algorithm we acquire new method with following favorable circumstances: it is exact enough; it takes into record time intervals between client activities; it gives reasonable for a human master interpretation of produced behavior models as "Assuming… THEN" rules. By modeling the temporal examples of fraudsters, CopyCatch detects ill-conceived Page Likes on Facebook. Through cautious design and usage, CopyCatch is appropriated on Hadoop, keeps running on Facebook's 10-billion edge Page Like graph, and detects both fake and hijacked accounts. To make recommender frameworks interpretable, we design a brief added substance co-grouping model of client behavior. ACCAMS matches the exactness of best in class methods, while having a 4 times smaller model.

References:

1. **Liang Zhao, Zhikui Chen, Yueming Hu, Geyong Min and Zhaohua Jiang**, “Distributed Feature Selection for Efficient Economic Big Data Analysis”, JOURNAL OF LATEX CLASS FILES, VOL. 13, NO. 9, SEPTEMBER 2014.

2. **Mohammad Shorfuzzaman**, “LEVERAGING CLOUD BASED BIG DATA ANALYTICS IN KNOWLEDGE MANAGEMENT FOR ENHANCED DECISION MAKING IN ORGANIZATIONS”, International Journal of Distributed and Parallel Systems (IJDPS) Vol.8, No.1, January 2017.

3. **Dr. Venkatesh Naganathan**, “Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends”, 2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis.

4. **Rakesh Ranjan Kumar & Binita Kumari**, “Visualizing Big Data Mining: Challenges, Problems and Opportunities”,) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3933-3937.

5. **Vivekananth.P , Leo John Baptist.A**, "An Analysis of Big Data Analytics Techniques", Volume-5, Issue-5, October-2015 International Journal of Engineering and Management Research Page Number: 17-19.

6. **Dr.M.Padmavalli**, "Big Data: Emerging Challenges of Big Data and Techniques for Handling", IOSR Journal of Computer Engineering (IOSR-JCE).

7. **Althaf Rahaman.Sk,Sai Rajesh.K,.Girija RaniK**, "Challenging tools on Research Issues in Big Data Analytics", International Journal of Engineering Development and Research.

8. **Sofiya Mujawar , Aishwarya Joshi**, "Data Analytics Types, Tools and their Comparison", International Journal of Advanced Research in Computer and Communication Engineering.

9. **Yolanda Gil**, "Teaching Big Data Analytics Skills with Intelligent Workflow Systems", National Conference of the Association for the Advancement of Artificial Intelligence (AAAI), Phoenix AZ, 2016.

10. **Revanth Sonnati**, "Improving Healthcare Using Big Data Analytics", NTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 6, ISSUE 03, MARCH 2017.

11. **Mengru Li, Hong Fu , Ruodan Sun and Che Che**, "The Application of Big Data Analysis Techniques and Tools in Intelligence Research", International Conference on Communications, Information Management and Network Security (CIMNS 2016).

12. **M. Dhavapriya, N. Yasodha**, "Big Data Analytics: Challenges and Solutions Using Hadoop, Map Reduce and Big Table", International Journal of Computer Science Trends and Technology (IJCST) – Volume 4 Issue 1, Jan - Feb 2016.

13. **Manisha R. Thakare, S. W. Mohod and A. N. Thakare**, "Various Data-Mining Techniques for Big Data", International Conference on Quality Up-gradation in Engineering, Science and Technology (ICQUEST2015).

14. **M.Chalapathi Rao, A.Kiran Kumar**, "Challenges arise of Privacy Preserving Big Data Mining Techniques", International Research Journal of Engineering and Technology (IRJET).

15. **B R Prakash , Dr. M. Hanumanthappa**, "Issues and Challenges in the Era of Big Data Mining", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).