The Seybold Report

ISSN 1533-9211

DOI 10.17605/OSF.IO/96NDA

LINK PREDICTION ON MULTI ATTRIBUTE EXTRACTED SOCIAL NETWORK USING MACHINE LEARNING MODELS

UGRANADA CHANNABASAVA[1] and B K RAGHAVENDRA2

[1]Department of Information Science & Engineering, Don Bosco Institute of Technology, Bengaluru, Karnataka, Affiliated to Visvesvaraya Technological University, Karnataka, India. Email: channasan11@gmail.com
[2]Professor and Head, Department of Information Science & Engineering, Don Bosco Institute of Technology, Bengaluru, Karnataka, Affiliated to Visvesvaraya Technilogical University, Karnataka, India.
Email: raghavendra.bk69@gmail.com

Abstract

An effective ensemble-based consensus-based multi-feature learned social media link prediction model is created in this research. Contrary to traditional methodologies, an improvement paradigm with many levels was taken into account. Where the initial emphasis was on extracting the most characteristics feasible that showed inter-node relationships for high prediction accuracy. We extracted local, Behavioural, as well as topological features, such as the Jaccard coefficient, cosine similarity, number of followers, intermediate followers, ADAR, shortest path, page rank, Katz coefficient index, hitting time of hops, and preferential attachment, taking into account the robustness of the various feature sets. The suggested link-prediction model was reinforced by using all of these attributes as link-signifiers, allowing it to be trained over larger datasets and with greater accuracy. Undoubtedly, using the aforementioned multiple features-based strategy might result in more accuracy and dependability, but at the expense of more computation. Different feature selection techniques, including the Gini index (GI), information gain (IG), PCA, and cross correlation (CC), were used to prevent it. These feature selection techniques were used with two goals in mind: first, to determine which types of features can have greater accuracy, and second, to minimise unnecessary computation. According to this study, cosine similarity-based characteristics don't significantly affect final categorization. In order to categorise each node-pair as Linked or Not-Linked, we developed a unique consensus-based ensemble learning model employing deep-neuro computing methods (ANN-LM with several hidden layers). Our suggested link-prediction model outperformed existing machine learning techniques in terms of link-prediction accuracy (98.7%), precision (0.95), recall (0.99), and F-Measure (0.93).

Keywords: Ensemble Learning Model, Feature Extraction, Social Network, Feature Selection

1. INTRODUCTION

The rapid development of internet and software computer technologies has opened up new possibilities for mass-market applications. Social media networks include Myspace, YouTube, LinkedIn, Facebook, Twitter, and Instagram and others have drawn the most attention among the major apps. According to a recent report, using social media is one of the most common internet activities and has been growing rapidly.

2. RELATED WORK

In directed weighted graphs, the work principle ensemble framework was introduced. It uses phases including community formation, community optimization, probabilistic network embedding, and classifier prediction. To divide the training graph into several communities, they employed edge betweeness and modularity maximisation. These communities were then

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

optimised using the cluster centroid approach. The link prediction issue in an undirected graph can be solved using the suggested method [1].

With the use of various combinations of meta-features, the suggested EML (Ensemble of ML-KNN) technique may automatically suggest a variety of suitable algorithms for various classification problems. By adjusting feature sets, an ensemble of ML-KNN may be created, increasing variety and enhancing recommendation performance. Emphasis on further enhancing the effectiveness and quality of EML as a starting point for ML-KNN, which may be used to locate more appropriate neighbours and enhance recommendation performance [2].

Numerous firms use this framework to suggest relevant positions to job searchers. Additionally, it handles the problems associated with data overload. The success of the recommender system across a vast quantity of data has helped it gain popularity in recent years. Using information from users of social networks, online recruitment systems, one of its application areas, employ recommender algorithms to match users with relevant jobs. Additionally, a geo-area-based recommender system is implemented, which enables suggested candidates to find the specific location of the companies [3].

User participation in numerous online social networks began. For cross-platform recommendation, information sharing, etc., the capacity to recognise the same individual across social networks may be a considerable value. For predicting anchor connections, authors proposed a number of consistency-based methods (MC). It iteratively uses both intra-layer structure information from network representation learning and inter-layer structure information. Utilizing the intra-layer structural information, a matrix factorization-based network representation learning method is employed to generate embedding vectors that capture the global structural characteristics of nodes. Then, a radial basis neural network's mapping function is trained to map embedding vectors from several spaces to a single space. The anchor links among node pairs are lastly predicted by accounting for both the interlayer and intra-layer structures [4].

According to the authors, the topic of link prediction in organisational social networks based on email correspondence between workers has been explored. The task of finding communities on the network may be aided by the link prediction. On the basis of how closely the vertices are spaced from one another, many similarity measures have been investigated. One was able to choose similarity metrics that might be used in the link prediction process in order to properly assign vertices (workers) to communities (units) in organisational networks (organization). The experiment's results were compared to genuine public organisation structure. [5].

Both homogeneous and heterogeneous procedures were used in the generation of the studied ensembles, however researchers are urged to evaluate the effects of using bigger ensembles and other heterogeneous ensembles as the diversity criteria may be achieved by combining several ML techniques. In addition, choosing the members of the ensemble base learners is a difficult procedure that has to be solved by more future contributions. The ensemble models were built using a large number of single learners. Only three kinds, Regression Trees, ANN, and AR, have been extensively examined in the publications that have been evaluated [6].

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

By examining directed sub-graphs from the various Facebook commotion networks like, comment, share, and post, it is possible to see how links arise within these networks and how sub-graphs transition into one another. New ideas, such as the directed sub-graph transitions matrix and associated Hasse diagram, were proposed in order to better evaluate sub-graph transitions. The backdrop of sub-graph transitions an innovative method of link prediction, which forecasts links in the network, is investigated. We find that none of the transitions are adequate for link prediction. Author utilised 10 out of 53 transitions to expedite the process and increase link prediction accuracy. The results of the studies show that the suggested strategy works better than existing link prediction methods [7].

In order to forecast missing links, the author has addressed certain important social network analysis elements such group norms, information dissemination, and various channels of contact. In order to take these considerations into account, a fuzzy-based link prediction model FLP-ID that takes into account community relevance in the multiplex network is provided. Additionally, covered in this paper are FLP-framework, ID's algorithm, and fuzzy network analysis. Additionally, the framework was evaluated and compared by the author against numerous real-world network datasets and cutting-edge algorithms [8].

The author used a model for predicting popularity based on sociological theory to address the issue that the forecasting accuracy of existing approaches is insufficient. He discovers a strong linear relationship between the percentage of devoted followers on the Facebook homepage and frequent shares in the initial and subsequent popularity. The statistical findings about Facebook serve as a reminder that the social physics theory is crucial to the work of prediction. Additionally, an experimental research demonstrates the usefulness of the suggested approach. The experiment's findings show that the suggested model may perform better than the other models due to the use of the widely accepted exhaustion theory, which supports the model's efficacy in popularity prediction [9].

The author has analysed a corpus of 1361 papers and used an unsupervised machine learning technique called STM to automatically find latent topics within them. Each PDF contains postings made on the 502 official Facebook pages of Italian towns during the course of a single year from 2016 to 2018. The variables indicating the municipal costs per capita by function and a number of additional factors that might influence the prevalence of each issue are included in the STM estimate. The prevalence of each issue and the most pertinent costs per capita by function are found to be positively and significantly correlated by the model. This gives proof of how local governments strategically promote themselves on Facebook in an effort to increase their political legitimacy and public support [10].

The friend recommendation system uses several machine learning algorithms, such as the Random Forest Classifier, XGBoost, Light GBM, and Cat Boost, to suggest friends to social network users. The Random Forest and Light GBM have worse accuracy compared to XGBoost and CatBoost method, according to the author's comparison of the performance matrices of several machine learning techniques. Both the XG- Boost and CatBoost algorithms have an accuracy of 95% [11].

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

3. PROPOSED SYSTEM

This section mainly covers the implementation of the proposed system as a whole. Our suggested social network link-prediction methodology includes the Consensus based Ensemble Learning for Link-prediction because it is a multi-phased analytics challenge.

Data gathering: Designing a link-prediction model that can quickly forecast similar connections and learn across a large user base is essential given the complexity of social media and the needs of the modern world. As a result, 9437519 node-edges were taken from Stanford network analysis repository into account while collecting the entire set of users and drawing each network-graph. After getting the dataset, the various characteristics, including local, behavioural, and topological aspects, were obtained. Here are several approaches for choosing features.

a. cross-correlation(CC)

We used a Pearson correlation test on the input characteristics to conduct a cross-correlation test. While some feature sets were deleted, those with correlation coefficients greater than the level of significance (p=0.5) as described above were kept. The original structure kept the important feature components of each category while eliminating the unimportant feature elements from each feature type. The traits that were ultimately kept were then classified further in order to anticipate social media links. Gaining information (IG)

Typically, IG is defined as per the following equation (1). In a prediction issue, IG is utilised as a term-goodness criterion. By knowing if a word is present or absent in the data corpus or feature set, it may estimate the amount of bits required for link prediction.

$$\mathbf{IG(t) = -\sum_{i} Pr(c_i) log\, Pr(c_i) + Pr(t) \sum_{i} Pr(c_i|t)\, log\, Pr(c_i|t) + Pr(t) \sum_{i} Pr(c_i|t)\, logPr(c_i|t)} \quad \textbf{(1)}$$

b. Gini Index

Gini Index often measures the impurity as per (2).

$$\mathbf{m(s) = \sum_{i \neq j} \widehat{P_{si}}\, \widehat{P_{sj}} = 1 - \sum_{j} \widehat{P_{sj}}} \quad \textbf{(2)}$$

The variance impurity is generalised by its functionality (signifying the variance of a distribution related with the two classes I and j). The predicted error rate when the class label is randomly selected from the feature distribution at a node is another way to put it. For the current two-class classification issue, our impurity criteria is more peaked at the same probability as the traditional entropy-based approaches. Because of this, GI-based feature selection is a good (candidate) solution for our feature selection issue. One may describe the

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

GI index as an alternative to the IG measure. The chance of a feature variable remaining in a feature subset is calculated mathematically using (3).

$$\mathbf{Gini(S) = 1 - \sum_{i=1,\ldots,m} P_i^2} \qquad \mathbf{(3)}$$

Pi, which is calculated as |Ci,S|/|S| in (3), indicates the chance that a tuple in feature set S belongs to class Ci. Notably, it would be summed over two (m=2) classes because our task is a two-class classification issue (appropriate for link-prediction or unsuitable for link prediction). We projected chosen features by each feature selection method separately since our proposed link-prediction model was designed to assess the classification performance of each feature set (after feature selection procedure). In addition, we combined every feature into a single feature vector we dubbed All-Matrix (AM). For additional two-class classification, these feature vectors or feature sets were provided as input to the classifiers. The next sections provide a thorough explanation of the proposed ensemble learning-based categorization model.

3.1 Consensus based Ensemble for Link-Prediction

We have used machine learning methods since our suggested link-prediction issue denotes a two-class classification problem. Generalizing the outcome by one machine learning algorithm is suspect and dubious given that multiple machine learning methods display varying performance over the same dataset. In this research, we suggested a consensus-based ensemble learning (CEL) method for social network link prediction to address this issue. As opposed to conventional machine learning methods, our proposed CEL incorporates a number of base classifiers from several operational philosophies. We employed six distinct base-classifiers for the CEL solution, which are listed below.

$\mathbf{CEL_{BC1}}$: ANN LM with 1 hidden layer

$\mathbf{CEL_{BC2}}$: ANN LM with 2 hidden layers

$\mathbf{CEL_{BC3}}$: ANN LM with 3 hidden layers

$\mathbf{CEL_{BC4}}$: ANN LM with 4 hidden layers

$\mathbf{CEL_{BC5}}$: ANN LM with 5 hidden layers

$\mathbf{CEL_{BC6}}$: SVM

Notably, we achieved learning over "deep-features" in our proposed classification model using ANN versions of Levenberg Marquardt (LM) learning approaches with multiple hidden layers. It is hypothesised that classification accuracy may occasionally be impacted by a rise in the buried layer (however at the cost of increased computation). We used ANN-LM variants with several hidden layers as the foundation classifiers in this case since we wanted to increase classification accuracy. Being a We implement the aforementioned base classifiers in the consensus-based ensemble learning model (CEL) in such a way that each classifier predicts the link between peer nodes in a unique fashion. The consensus model forecasts each user-pair as Linked or Not-Linked when we finally receive an individual prediction result (Linked or Not-

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

Linked). In this case, the final link-prediction for each pair of participating nodes is computed by the consensus model using the maximum voting ensemble, also known as the maximum voting criteria.

The following is a sample of the machine learning algorithms used as the basis classifier.

a. Deep Neuro-Computing

One of the most popular machine learning techniques for data learning and classification is the neural network, often known as an artificial neural network (ANN). Although ANN has gone through many stages of evolution depending on computational complexity and adaptive computation, its resilience makes it effective for usage in a variety of classification situations. The performance of ANN is closely correlated with the associated learning technique, according to an in-depth investigation. As a result, ANN has developed into other forms, such as ANN with steepest gradient (SD), ANN with gradient descent (GD), ANN with RBF (ANN-RBF), ANN with Levenberg Marquardt (ANN-LM), etc., depending on the learning technique used. However, ANN-LM and ANN-GD have been proven to be more successful when used in conjunction with non-linear heterogenous data categorization. Even with a large non-linear feature set, ANN-GD avoids local minima and convergence problems, unlike ANN-SD. Similar to how ANN-LM is more robust than ANN-SD and ANN-GD alone. Additionally, ANN-LM may be set up to have both ANN-SD and ANN-GD features, which improves performance stability even with big, non-linear, and heterogeneous data. In light of this, we used ANN-GD and ANN-LM as basis classifiers in this research. However, in order to build ANN-GD and ANN-LM with various hidden layers, we created a deep-neuro-computing (DNC) environment after discovering the importance of "deep-features" to any classification task. With 1, 2... 5 hidden layers, we created ANN-GD and ANN-LM. This

$$\mathbf{O_h = \frac{1}{1 + e^{-I_h}}} \qquad \textbf{(4)}$$

I_h Stands for the input at the hidden layer in (4). ANN is frequently described as Y'=f(W,X), where Y' denotes the output vector and X and W, respectively, denote the allied input and weight values In order to attain greater accuracy, ANN functionally employs certain error functions like mean square error (MSE), which is calculated using (4).

$$\mathbf{MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i' - y_i)^2} \qquad \textbf{(5)}$$

In above equation (5), y presents the observed output value, while the expected value is y_i'. As stated above, the key difference between the different ANN variants is the way it schedules or updates its weight values over training. A snippet of the different ANN variants (i.e., ANN-GD and ANN-LM) is given as follows.

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

b. ANN-GD

Ten different ANN classification environments were generated by the DNC setup. Notably, the primary goal of the DNC idea for 10 distinct ANN variants was to determine if the performance of the deep features retrieved by ANN variants with greater hidden layers could be improved. However, it is posited that ANNs with more hidden layers will perform better. Since ANN-GD and ANN-LM have been used as the primary base classifier variations in our suggested DNC environment, a sample for the same

In order to learn from specific input data or patterns, ANN functionally imitates human brain capabilities. As a result, by learning from such input patterns, it organises unknown information into desired categories. Input, hidden, and output layers are the three layers that make up an ANN. If we look at the architecture of ANN, we can see that it consists of several neurons that represent the input data that will be processed further at various intermediary levels (such hidden layers) for classification (at the output layer). An artificial neural network (ANN) uses the error-reduction approach to learn over the input data, estimating the discrepancy between the predicted and actual outputs as it goes (signifying error). Up until the error output is zero or close to zero, the learning process is ongoing. The final output is therefore anticipated to achieve zero-error outputs at the output layer. An ANN is anticipated to conduct two-class classification at the output layer in light of the current link-prediction issue.

Our suggested model feeds the ANN with the various characteristics that were retrieved for each user, varying the number of hidden layers (here, 1, 2, 3, 4 and 5 hidden layers are used with both ANN-GD as well as ANN-LM). The linear activation function of the ANN is used at the input layer to produce output that is identical to the input (i.e., O o=I i), while the output of the hidden layer is given to the input of the output layer. The ANN's output layer, in particular, uses the sigmoid function (4) to produce O_h

Let the regression for the learning method, while reducing error value be (6).

$$\mathbf{w}^* = \underset{\mathbf{w}}{\mathbf{argmin}}\, \mathbf{L}(\mathbf{w}) \tag{6}$$

$$\mathbf{L}(\mathbf{w}) = \sum_{\mathbf{t=1}}^{\mathbf{N}} \mathbf{L}\left(\mathbf{y_t}, \mathbf{f_w}(\mathbf{x_t})\right) + \boldsymbol{\lambda}\mathbf{R}(\mathbf{w}) \tag{7}$$

In ANN-GD setup, $\mathbf{f_w}(\mathbf{x})$ factor states the non-linear weight$\mathbf{w}$, and thus it intends to achieve a local optimum for (7) using GD method, which updates $\mathbf{w}$ iteratively by updating $\mathbf{w_t}$by$\mathbf{w_{t+1}}$.

$$\mathbf{w_{t+1}} = \mathbf{w_t} - \boldsymbol{\eta_t}\boldsymbol{\nabla}\mathbf{L} \tag{8}$$

$$\mathbf{w_{j,t+1}} = \mathbf{w_{j,t}} - \boldsymbol{\eta_t}\frac{\partial \mathbf{L}}{\partial \mathbf{w_j}} \tag{9}$$

In (8), the parameter $\boldsymbol{\nabla}\mathbf{L}$signifies the error value, which is mathematically given as (10).

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i' - y_i)^2 \quad (10)$$

The learning rate is stated in (9) as η_t, which decreases over time t. Each node pair is therefore classified as linked or Not-Linked by conducting GD-based weight estimate and learning it. Notably, we used ANN-GD in our proposed study with 1, 2, 3, 4, and 5 hidden layers, where each configuration served as a separate base classifier.

c. ANN-LM

ANN-LM has more resilience in learning across huge non-linear data input than traditional ANN-GD and ANN-SD. Sum of Squares (SoS), a multivariate function that measures the least value of non-linear real-valued functions, is localised by ANN-LM. With the help of this capability, ANN-LM can change weights more quickly and effectively. Additionally, it stays clear of local minima and other convergence-related problems, making it appropriate for big datasets. As previously said, ANN-LM embodies the abilities of both ANN-SD and ANN-GD, which are chosen via adaptive learning rate selection, aiding retrieval and speedy error reduction. When learning, ANN-LM applies (11) to update weights.

$$\mathbf{W_{j+1}} = \mathbf{W_j} - \left(\mathbf{J_j^T J_j} + \mu\mathbf{I}\right)^{-1}\mathbf{J_j e_j} \quad (11)$$

In (11), the parameter $\mathbf{W_j}$ signifies the at-hand weight while $\mathbf{W_{j+1}}$ presents the updated weight. Similarly, $\mathbf{I}$ indicates the identity matrix, while the Jacobian matrix is given by $\mathbf{J}$ (12). Here (11), μ states Lower values of and the combination coefficient cause ANN-LM to respond in an ANN-GD manner, whereas larger values require it to behave in an ANN-SD manner.

$$\mathbf{J} = \begin{bmatrix} \frac{\mathbf{d}}{\mathbf{dW_1}}(\mathbf{E_{1,1}}) & \frac{\mathbf{d}}{\mathbf{dW_2}}(\mathbf{E_{1,1}}) & \cdots & \frac{\mathbf{d}}{\mathbf{dW_N}}(\mathbf{E_{1,1}}) \\ \frac{\mathbf{d}}{\mathbf{dW_1}}(\mathbf{E_{1,2}}) & \frac{\mathbf{d}}{\mathbf{dW_2}}(\mathbf{E_{1,2}}) & \cdots & \frac{\mathbf{d}}{\mathbf{dW_N}}(\mathbf{E_{1,2}}) \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\mathbf{d}}{\mathbf{dW_1}}(\mathbf{E_{P,M}}) & \frac{\mathbf{d}}{\mathbf{dW_2}}(\mathbf{E_{P,M}}) & \cdots & \frac{\mathbf{d}}{\mathbf{dW_N}}(\mathbf{E_{P,M}}) \end{bmatrix} \quad (12)$$

N stands for the total weight counts in the equation above, and P shows the input features. The result is provided by M.

d. Consensus based Ensemble Learning

The machine learning algorithms that were previously covered were used as the basic classifier to carry out two-class classification. Every node pair was assigned either a Linked ("1") or Not-Linked ("0") label using the classifier. Thus, in order to estimate consensus-based prediction, our suggested consensus-based ensemble learning (CEL) model employed the idea of

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

maximum voting ensemble (MVE) to acquire the classified label for each node-pair. The final prediction result in this method was determined by taking the highest prediction output (1 or 0) from each base classifier. In other words, our suggested CEL model predicts a node pair as "Linked" if a total of 7 base classifiers predict it to be "Linked" whereas 5 other base classifiers predict it to be "Not-Linked". In order to classify each node pair as connected or not linked, our suggested CEL model obtains the consensus of prediction by each base classifier. Thus, this work successfully solved the overall connected prediction issue. The next part provides a thorough explanation of the simulation's findings and related implications.

4. RESULTS AND DISCUSSION

For the purpose of predicting social media links, we created a unique and reliable multi-feature ensemble learning model in this study. Contrary to the majority of existing methodologies, we used a multi-phase enhancement-based idea where the main goal was to accurately anticipate social media links by utilising the most features and increased machine learning capabilities while assuring minimal computation. We employed a social media dataset containing users and related edge-mapped information that was made accessible to the public in order to simulate a real-world application situation with an enormously vast user base. There were a total of 1862220 people and 9437519 edges in the input data under consideration, where the edge value denoted the association value discovered by crawling over the online nodes. As a result, getting the nodes and their associated edge values was described as a two-class classification task, with the primary goal being to determine whether or not the two unknown users are related. In order to do it, we concentrated on extracting as many characteristics as we could in this study. Although writers have included either temporal, local, behavioural, or topological characteristics in the majority of current approaches, the majority of these attempts only use one feature for link-prediction, which does not assure their supremacy to provide optimal link-prediction. Because standalone feature-based prediction can't take into account other types of features, it is unable to comprehend the latent data that underlies social media usage, preferences, behaviours, etc. As a result, this may produce inaccurate forecast results. We identified multi-trait information, including temporal, behavioural, and topological, to address this issue. We acquired a variety of characteristics, such as the Jaccard coefficient, cosine similarity, ADAR, Page Rank, common neighbor, preferred attachment, shortest distance, as well as the page rank and preferential attachment, in order to do this. Therefore, combining all of these attributes allowed for the retrieval of a sizably large collection of features for the best link-prediction. It should be mentioned that we tested each feature's effectiveness for link-prediction purposes. The cosine-similarity was the only variable that had a meaningful influence on link prediction, according to the correlation analysis with a 95% confidence interval (CI), also known as the level of significance with a 95% CI. The characteristics having the strongest association to link-prediction probability were the Jaccard coefficient, number of followers, inter-node distance, ADAR, common neighbor, and hops. The comprehensive result graphs for the various characteristics and their influence on prediction probability are not included in this publication due to memory or space limitations.

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

Since we needed 1862220 user records with related edge data, extracting the relevant features required a massive calculation on a general-purpose computer system (here, we used RAM of 8 GB and an Intel i3 CPU), therefore we only used 1 lakh user records for feature extraction. To put it another way, out of the 1862220 users, we extracted characteristics for a total of 1 Lakh people, which were then used in feature selection and classification processing. Unquestionably, the multi-trait characteristics for 1 Lakh people were too large to compute on the standard system, therefore we used the feature selection approach to pick just the most important features to employ in the remaining computation. We used a variety of feature selection techniques, including the rank sum test, PCA, cross-correlation, Information Gain, and Gini Index, to decrease computational cost. Here, our main goal was to create a minimally complex set of characteristics that could produce the best prediction accuracy with the least amount of processing. Additionally, we sought to pinpoint the most effective feature selection strategy for link-prediction, particularly while considering a sizable user population and allied feature size. As a result, we input the features to the consensus-based ensemble learning (CEL) model after getting them from each feature selection process. Our suggested CEL model included classifiers from many machine learning categories, including regression, decision trees, pattern mining, and neuro-computing; as a result, it may be referred to as a heterogeneous ensemble learning model. We used logistic regression, decision trees (C5.0), SVMs with polynomial kernels, and ANN variations as our basic classifiers. We utilised ANN-GD with various hidden layers, taking into account the advantage of ANN-GD over traditional ANN-SD for non-linear big feature learning (here, we applied ANN-GD with 1, 2, 3, 4 and 5 hidden layers, constituting a deep-neuro-computing environment). To create a deep-neuro computing environment, we similarly employed ANN-LM with several hidden layers. In order to conduct social media link prediction, a total of 6 base-classifiers (in addition to the suggested CEL-MVE ensemble classifier) were implemented. We used the maximum voting concept, also known as MVE, as a consensus-based learning and prediction method to forecast the likelihood of an inter-node or inter-user link. Notably, each base classifier assigns a "1" or a "0" depending on whether it believes each node-pair to be linked or not. As a result, the CEL-MVE model predicts whether each node pair is related or not based on the labels assigned to each node pair. We obtained confusion matrix factors to assess performance. We measured the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values in order to achieve it. We arrived at the performance metrics accuracy, precision, recall, and F-Measure using the aforementioned matrix values, as shown in Table I.

a. Feature Sensitiveness Analysis

In light of the foregoing discussion, we used a variety of feature selection techniques on the retrieved features, including the rank sum test, cross correlation, PCA, Gini Index (GI), and Information Gain (IG), in order to keep only those features that were significant and could guarantee the best performance. First, we evaluated the appropriateness of the feature selection approach for the final classification before looking at the performance of the suggested social network link-prediction model. In order to do so, we acquired Accuracy, precision, recall and F-Measure values for the various feature sets. For various characteristics using machine

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

learning models Table II, Table III, Table IV, Table V contemporary the accuracy, precision, recall and f-score outputs using various machine learning models for the various characteristics.

Table I: Performance Parameters

Parameter	Mathematical Expression	Definition
Accuracy	$\frac{(TN + TP)}{(TN + FN + FP + TP)}$	Represents the percentage of modules that are examined that are projected to be fault-prone.
Precision	$\frac{TP}{(TP + FP)}$	Indicate the extent to which the same findings are obtained from repeated measurements under the same conditions.
Recall	$TP/((TP + FN))$	It shows how many of the pertinent things need to be found.
F-measure	$2.(Recall.Precision)/(Recall + Precision)$	It combines the recall and accuracy numerical values to produce a single score that is defined as the harmonic mean of the two.

Table II: Accuracy Performance

Existing system							Proposed system							
Feature selection	Machine Learning (Base Classifier)					Ensemble	Feature selection	Machine Learning (Base Classifier)						Ensemble
Techniques	LOGR	DT	ANN-GD-1H	ANN-GD-2H	SVM	CEL-MVE	Techniques	ANNLM-1H	ANNLM-2H	ANNLM-3H	ANNLM-4H	ANNLM-5H	SVM	CEL-MVE
AM	91.9	95.5	86.0	88.5	90.8	**97.1**	GI	95.0	95.3	95.1	95.1	94.9	90.8	**97.4**
RST	89.5	96.2	87.1	89.2	92.4	**98.4**	IG	88.1	88	88.2	88.1	87.9	78.7	**98.7**
CC	92.2	95.4	81.9	87.2	86.5	**95.1**	CC	94.9	95.4	95.5	**95.6**	95.7	86.5	95.1
PCA	78.9	81.1	81.7	86.4	84.5	**96.1**	PCA	89.1	89.1	92.3	91.2	91.4	84.5	**96.8**

Table III: Precision Performance

Existing system							Proposed system							
Feature selection	Machine Learning (Base Classifier)					Ensemble	Feature selection	Machine Learning (Base Classifier)						Ensemble
Techniques	LOGR	DT	ANN-GD-1H	ANN-GD-2H	SVM	CEL-MVE	Techniques	ANNLM-1H	ANNLM-2H	ANNLM-3H	ANNLM-4H	ANNLM-5H	SVM	CEL-MVE
AM	0.91	0.95	0.79	0.81	0.91	**0.93**	GI	0.91	0.92	0.92	0.92	0.91	0.83	**0.95**
RST	0.84	0.95	0.80	0.83	0.88	**0.93**	IG	0.82	0.82	0.82	0.82	0.82	0.79	**0.83**
CC	0.88	0.94	0.73	0.91	0.79	**0.93**	CC	0.91	0.92	0.92	0.92	0.93	0.79	**0.95**
PCA	0.76	0.87	0.81	0.83	0.86	**0.85**	PCA	0.83	0.83	0.87	0.87	**0.88**	0.86	0.85

Table IV: Recall Performance

Existing system							Proposed system							
Feature selection	Machine Learning (Base Classifier)					Ensemble	Feature selection	Machine Learning (Base Classifier)						Ensemble
Techniques	LOGR	DT	ANN-GD-1H	ANN-GD-2H	SVM	CEL-MVE	Techniques	ANNLM-1H	ANNLM-2H	ANNLM-3H	ANNLM-4H	ANNLM-5H	SVM	CEL-MVE
AM	0.93	0.95	0.98	0.97	0.97	**0.98**	GI	0.99	0.99	0.98	**0.99**	**0.99**	**0.99**	0.99
RST	0.96	0.97	0.98	0.98	0.98	**0.98**	IG	0.97	0.97	0.97	**0.98**	0.97	0.78	0.97
CC	0.97	0.96	0.97	0.97	0.98	**0.98**	CC	**0.99**	0.98	0.98	**0.99**	0/98	0.98	**0.99**
PCA	0.83	0.89	0.83	0.90	0.82	**0.98**	PCA	0.97	0.97	**0.98**	0.96	0.95	0.82	**0.98**

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

Table V: F-Measure Performance

Existing system							Proposed system							
Feature selection	Machine Learning (Base Classifier)					Ensemble	Feature selection	Machine Learning (Base Classifier)						Ensemble
Techniques	LOGR	DT	ANN-GD-1H	ANN-GD-2H	SVM	CEL-MVE	Techniques	ANNLM-1H	ANNLM-2H	ANNLM-3H	ANNLM-4H	ANNLM-5H	SVM	CEL-MVE
AM	0.92	0.95	0.90	0.91	0.90	**0.97**	GI	0.91	0.92	0.92	0.92	0.93	0.89	**0.93**
RST	0.91	0.94	0.87	0.90	0.91	**0.97**	IG	0.87	0.88	0.88	0.88	0.89	0.74	0.89
CC	0.92	0.92	0.89	0.89	0.88	**0.97**	CC	0.91	0.91	0.92	0.92	**0.93**	**0.93**	**0.93**
PCA	0.83	0.85	0.77	0.74	0.81	**0.89**	PCA	0.88	0.88	0.89	0.89	0.90	0.89	**0.91**

As can be seen from the results (Table II), our suggested CEL-MVE model achieved the greatest accuracy of 97.4% with the Gini Index (GI) feature, while the maximum accuracy of 98.7% was obtained with the Information Gain (IG) based feature. Comparably, PCA-based features may attain CEL-maximum MVE's accuracy of 96.8%. When compared to the accuracy of the previous system, which was 95.1% for cross correlation, 96.1% for PCA, 98.4% for RST, and 97.1 for AM, cross correlation (CC) features demonstrated the highest accuracy of 95.1% using ANN-LM (with 4 hidden layers). Table III shows how well each algorithm performs in terms of accuracy. With regard to the findings, it can be shown that our suggested CEL-MVE classifier model with CC and AM has maximum precision of 0.95, PCA by 0.85, and RST by 0.83 compared with the current system's 0.93 for AM, RST, and CC, and 0.85 for PCA. Table IV shows the recall performance for several machine learning methods using various features. Recall of 0.99, an unquestionably important performance metric, is displayed by our suggested CEL-MVE ensemble classifier with AM and CC as features. It confirms the effectiveness and reliability of our suggested link prediction methodology. Additionally, when comparing the overall results for recall, ANN-LM with a greater number of hidden layers outperformed other base-classifiers. According to this, ANN-LM may be thought of as a possible machine learning classifier for the link-prediction problem. Compared to previous recall values of 0.98 for AM, CC, PCA, and RST, CEL-MVE and ANNLM-4H had a recall of 0.99. Table V shows the F-Measure performance for various (chosen) features and machine learning algorithms. The features GI and PCA both obtained maximum F-Measures of 0.89 and 0.93 for CEL-MVE and ANN-LM with five hidden layers. The robustness of ANN-LM with more hidden layers is undeniably demonstrated.

The overall findings (Tables II to V) show that GI, IG, and CC characteristics have a larger capacity to provide improved accuracy and trustworthy link-prediction. Similar to this, the relative performance evaluation shows that the suggested CEL-MAE model continues to have greater accuracy (98.7%), precision (0.95), recall (0.99), and f-score (0.93), indicating superior performance over cutting-edge base classifiers. As a consensus-based link-prediction, CEL-MVE has a greater and more acceptable reliability than a single classifier.

b. Machine Learning Performance Assessment

Although the results previously mentioned confirmed the suggested model's effectiveness, the various features also indicated the adequacy of the features and the accompanying machine

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

learning model. However, we did 5-fold cross validation and acquired results in the form of accuracy, precision, recall, and F-Measure to evaluate the final performance of our suggested consensus based multi-feature learned ensemble learning model for link-prediction. As seen in our simulation, we were able to extract the features for a total of 1 lakh users. However, training over such enormously large features or models was laborious. Due to this, we trained over 5000 people to assess the effectiveness of our suggested link-prediction. 5-fold cross validation is carried out. The suggested CEL-MVE model performs more consistently, and as a result, must be regarded as the best alternative, according to observations of overall performance and relative significances. The applicability and dependability of CEL-MVE classifiers are more reasonable and suggestible, even if ANN-LM with higher hidden layers also function adequately. As a result, we are able to confirm in this study that the CEL-MVE model combined with IG characteristics is effective for predicting future social media links.

c. Machine Learning Performance Assessment

Undoubtedly, our suggested consensus-based multi-feature learned ensemble model performed better than the traditional state-of-the-art machine learning models (as a standalone classifier); nonetheless, we used a qualitative way to compare performance to other current approaches. In this way, we contrasted the effectiveness of our suggested algorithm with that of alternative techniques. Authors recently [12] created a prediction model for social network connection prediction that takes into account many variables such shared neighbours, admic/adler, followers, followees, etc. Notably, this strategy maintained the 98.4% accuracy of the pre-existing learning model while achieving a maximum accuracy of 98.7% with supervised learning method. The suggested CEL-MVE concept with GI feature set provides superior accuracy (98.7%) as well as acceptable recall measure performance when compared to the performance of our own proposed model. It confirms the robustness of our suggested model compared to earlier techniques. We have contrasted our suggested model with several alternative topological and behavioural feature-based methods to social media link prediction. Interestingly, our suggested strategy has outperformed other ways now in use.

5. CONCLUSION

The potential for business communities to use such platforms to reach their target audiences fast with the most information exchange has increased due to the exponential growth of internet technology and related applications, such as online social media. Even though social media has an enormously vast user base, it can be difficult to find interpersonal connections based on shared interests, preferences, or other factors. On the other hand, discovering many users who are as similar to one another as feasible can assist business communities in reaching a big percentage of people in a short amount of time with minimal overhead. Social media link prediction has become a promising strategy to achieve this goal. Although there have been a few studies on the social media evolution and link prediction tasks in the past, the bulk of the systems now in use either use a small range of user criteria, such as either topological or behavioural information. As a result, their effectiveness is restricted to precise link-prediction. Additionally, the majority of the approaches now in use use traditional machine learning

The Seybold Report
ISSN 1533-9211
DOI 10.17605/OSF.IO/96NDA

techniques as a stand-alone solution for link-prediction or related categorization. This research has placed a strong emphasis on using cutting-edge approaches using multi-feature learning and consensus-based ensemble learning principles to conduct social media link prediction in light of these important restrictions. In this study, many features including local features, topological features, and behavioural characteristics were retrieved from the node edge information in order to perform multi-feature learning. Thus, a significantly large but significant feature set was obtained, which could guarantee the highest level of prediction accuracy, using the various features such as the Jaccard coefficient, cosine similarity, number of followers, intermediate followers, ADAR, shortest path, page rank, Katz coefficient index, hitting time of hops, and preferential attachment for each node and allied node-pair. Different feature selection procedures, including the rank sum test, cross-correlation, principal component analysis, Gini index, and information gain were used as a result of the enormous feature volume and resulting influence on computational load. Notably, the evaluation of these feature selection techniques was carried out alone and in combination with feature sets, which ultimately assisted in determining the optimum feature selection technique for the relevant social media link prediction problem. In this paper, numerous algorithms, including logistic regression, decision trees, SVM, deep learning with ANN-GD and ANN-LM, and deep neuro-computing, were used to introduce diversity of performance by different learning concepts in recognition of the fact that different machine learning classifiers exhibit different performance over the same input data. Notably, ANN-GD and ANN-LM with five distinct hidden layers (1, 2, 3, 4, and 5 hidden layers) were used as the foundation classifiers as deep learning concepts. As a result, consensus-based ensemble learning was made possible by using a total of 6 basic classifiers. Maximum Voting Ensemble (MVE) theory was used to build consensus. It should be emphasised that in the proposed model, both the basic classifiers and the CEL (MVE) ensemble learning model underwent individual performance assessments. Consensus modelling allowed for more reliable categorization that was optimum. Being a two-class classification issue, the suggested link-prediction model allowed each base classifier and ensemble learning model to categorise node-pairs as Linked or Not-Linked. Each base classifier's prediction output was used by CEL, which operates based on MVE ideas, to conduct eventual classification or link-prediction. The base classifier ANN-LM with five hidden layers performs the maximum link-prediction accuracy (95.7%), recall (0.99), precision (0.93), and F-measure, whereas the suggested CEL-MVE model performs the highest cumulative accuracy of 98.7%, precision (0.95), recall (0.99), and F-measure of 0.93. (0.93). 1862220 people and 9437519 linked node-edges from Facebook social media data were used in this investigation. It affirms the robustness and applicability of the proposed model for actual social media link prediction purposes, which can assist business establishments in identifying or segmenting the connected users to spread their intended business information, service, or product information in order to gain a better and more competitive market share. Due to the fact that the suggested model only used traditional regression, decision trees, and machine learning techniques, additional cutting-edge algorithms like Extreme Learning Machine (ELM) and Least Square SVM (LSSVM) might be evaluated in the future for link-prediction.

ISSN 1533-9211

DOI 10.17605/OSF.IO/96NDA

Reference

1. Faima Abbasi," Exploiting optimised communities in directed weighted graphs for link Prediction", Online Social Networks and Media, 2022 https://doi.org/10.1016/j.osnem.2022.100222
2. Xiaoyan Zhu et.al, "Ensemble of ML-KNN for classification algorithm recommendation", Knowledge-Based Systems, 2021 https://doi.org/10.1016/j.knosys.2021.106933
3. Binny Parida et.al," Prediction of recommendations for employment utilizing machine learning procedures and geo-area based recommender framework", Sustainable Operations and Computers, 2022 https://doi.org/10.1016/j.susoc.2021.11.001
4. Yujie Yang," Anchor link prediction across social networks based on multiple Consistency", Knowledge-Based Systems, 2022 https://doi.org/10.1016/j.knosys.2022.109939
5. Paweł Szyman et.al," Link prediction in organizational social network based on e-mail communication ", 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022), 2022. https://doi.org/10.1016/j.procs.2022.09.463
6. Jianxing Zheng et.al," Attention-based explainable friend link prediction with heterogeneous context information", Information Sciences, 2022 https://doi.org/10.1016/j.ins.2022.03.010
7. Yingjie Liu et.al," Link prediction algorithm based on the initial information contribution of nodes", Information Sciences, 2022 https://doi.org/10.1016/j.ins.2022.07.030
8. Shashank Sheshar Singh," FLP-ID: Fuzzy-based link prediction in multiplex social networks using information diffusion perspective", Knowledge-Based Systems, 2022

   https://doi.org/10.1016/j.knosys.2022.108821
9. Xiaomeng Wang et.al," Predicting the security threats on the spreading of rumor, false information of Facebook content based on the principle of sociology", Computer Communications,2020 https://doi.org/10.1016/j.comcom.2019.11.042
10. Diego Ravenda et.al," The strategic usage of Facebook by local governments: A structural topic modelling analysis ", Information & Management, 2022 https://doi.org/10.1016/j.im.2022.103704
11. Ruksar Parveen et.al," Friend's recommendation on social media using different algorithms of machine learning", Global Transitions Proceedings, 2021 https://doi.org/10.1016/j.gltp.2021.08.012
12. Ugranada Channabasava et.al," Ensemble Assisted Multi-Feature Learnt Social Media Link Prediction Model Using Machine Learning Techniques, Revue d' Intelligence Artificielle, 2022 https://doi.org/10.18280/ria.360311