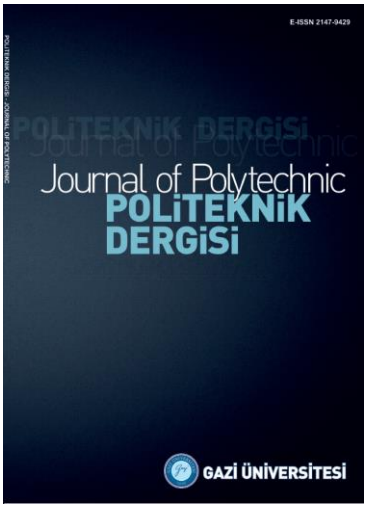

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)

URL: http://dergipark.org.tr/politeknik

Integration of algorithmic and local approaches for link prediction: An analysis on protein-protein interactions and social networks

Bağlantı tahmini için algoritmik ve yerel yaklaşımların entegrasyonu: protein-protein etkileşimleri ve sosyal ağlar üzerine bir analiz

***Yazar(lar) (Author(s))**: Hasibe CANDAN KADEM[1], Volkan ALTUNTAŞ[2]*

ORCID[1]: 0000-0001-5722-0811

ORCID[2]: 0000-0003-3144-8724

To cite to this article: Candan Kadem H. and Altuntaş V., "Integration of Algorithmic and Local Approaches for Link Prediction: An Analysis on Protein-Protein Interactions and Social Networks", ***Journal of Polytechnic***, *(*): *, (*).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Candan Kadem H. ve Altuntaş V., "Integration of Algorithmic and Local Approaches for Link Prediction: An Analysis on Protein-Protein Interactions and Social Networks", ***Politeknik Dergisi***, *(*): *, (*).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1563133

Integration of Algorithmic and Local Approaches for Link Prediction: An Analysis on Protein-Protein Interactions and Social Networks

Highlights

- *Studies on link prediction have been reviewed. In light of this, local metrics such as Adamic-Adar and Jaccard coefficient were used alongside global metrics to predict new links. By combining algorithmic and local approaches, new datasets on protein-protein interactions and social network datasets were derived. The success of link prediction with machine learning classification algorithms applied on the datasets was evaluated.*
- *The study brings a new perspective to the literature on link prediction for both protein-protein interactions and social network data.*

Graphical Abstract

In this study, Support Vector Machines (SVM) is used to predict the connectivity of protein-protein interaction (PPI) networks and social networks. This study aims for a new perspective by combining both local and global metrics in an algorithmic approach. Datasets consisting of Twitch, Facebook, Twitter and human gene PPI data are used. The data obtained were analyzed with similarity-based methods and the number of neighbors, Adamic-Adar index, Jaccard parameters and label values were calculated for each node and a new dataset was derived with these parameters. All datasets were evaluated using machine learning algorithms.

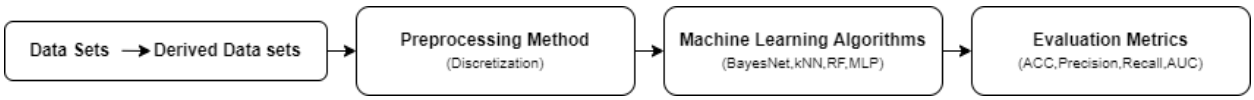

Figure. Model representation

Aim

The study combines algorithmic and local methods with machine learning techniques to improve link prediction in complicated networks. To increase forecast accuracy, we incorporate global and local indicators, in contrast to earlier approaches that only use one indication. To improve link prediction methods and get a better understanding of link formation processes, we intend to apply these techniques to social networks and protein-protein interaction networks in order to predict new connections and assess the effectiveness of these predictions across numerous datasets.

Design & Methodology

An SVM-based model was developed and used for link prediction. On four different datasets discretization was applied in the preprocessing stage of each dataset for link prediction. The performance of both models (with and without preprocessing) is evaluated using machine learning algorithms.

Originality

Creating new datasets using both local and global index values contributes to the field. In addition, it has been observed that the discretization method improves prediction success in social networks. The originality of the study is that link prediction is evaluated with different datasets and algorithms using a combination of global and local metrics and machine learning methods.

Findings

Especially the Twitch dataset showed the highest success. The application of discretization increased the performance values in all methods and datasets. Moreover, when the effect of the number of nodes and edges on the performance is analyzed, it is seen that the Twitch dataset with the highest number of edges has a superior performance in terms of link prediction.

Conclusion

The study demonstrates the effectiveness of machine learning algorithms for link prediction in complex networks. The discretization preprocessing technique improves the success of link prediction in social networks and plays an important role in the generation of new datasets.

Declaration of Ethical Standards

The authors of this article declare that the materials and methods used in this study do not require ethical committee permission and legal-special permission.

Integration of Algorithmic and Local Approaches for Link Prediction: An Analysis on Protein-Protein Interactions and Social Networks

Araştırma Makalesi / Research Article

Hasibe CANDAN KADEM[1*], Volkan ALTUNTAŞ[2]

[1,2] Faculty of Engineering, Department of Computer Engineering, Bursa Technical University, Turkiye

(Geliş/Received : 08.10.2024 ; Kabul/Accepted : 23.03.2025 ; Erken Görünüm/Early View : 12.04.2025)

ABSTRACT

Complex network analysis is applied in various fields such as network-based systems, social media recommendation systems, shopping platforms, and treatment methodologies. In this context, predicting the probability of connection between two nodes has become a focal point. Another significant aspect is the prediction of connections between proteins, especially with the increase in epidemic diseases. Link prediction methods, based on graph structures, aim to predict interactions between two nodes and measure the probability of connection between them. These methods proceed by relying on similarity values and can have multiple approaches, including local, global, and algorithmic. This study has emerged from a combination of algorithmic and local network approaches. Support Vector Machines are employed to predict connections in gene-protein networks and social network structures. Data sets from multiple social media platforms and human protein-protein interaction (PPI) data were utilized. Derived data were created by calculating local index values, including the number of neighbors, Adamic Adar index, Jaccard coefficient, and label values for each node. To enhance success rates, a model was developed that applied the discretization method as a preprocessing technique across all data sets. Machine learning algorithms such as Bayesian Networks, Multilayer Perceptron (MLP), Random Forest, and k-Nearest Neighborhood (kNN) were compared and evaluated. The results indicate that the Twitch dataset, which has the highest number of edges, produced successful outcomes. The contribution of edge numbers in the network structure to performance is highlighted, and it is observed that more successful metric values were obtained for the data with applied discretization.

Keywords: Link Prediction, Protein-Protein Interaction, Social Networks, Machine Learning, Biological Networks.

Bağlantı Tahmini için Algoritmik ve Yerel Yaklaşımların Entegrasyonu: Protein-Protein Etkileşimleri ve Sosyal Ağlar Üzerine Bir Analiz

ÖZ

Karmaşık ağ analizi, ağ tabanlı sistemler, sosyal medya öneri sistemleri, alışveriş platformları ve tedavi metodları gibi çeşitli alanlarda uygulanmaktadır. Bu bağlamda, iki düğüm arasındaki bağlantı olasılığını öngörmek odak noktası haline gelmiştir. Özellikle salgın hastalıklardaki artışla birlikte, proteinler arasındaki bağlantıların tahmin edilmesi önemli bir konudur. Graf yapılarına dayalı olan bağlantı tahmini yöntemleri, iki düğüm arasındaki etkileşimleri tahmin etmeyi ve bunlar arasındaki bağlantı olasılığını ölçmeyi amaçlar. Bu yöntemler, benzerlik değerlerine dayanarak ilerler ve yerel, global ve algoritmik gibi çeşitli yaklaşımlara sahip olabilir. Bu çalışma, algoritmik ve yerel ağ yaklaşımlarının bir kombinasyonundan ortaya çıkmıştır. Gen-protein ağları ve sosyal ağ yapılarında bağlantıları tahmin etmek için Destek Vektör Makineleri kullanılmıştır. Birden çok sosyal medya platformundan ve insan protein-protein etkileşimi (PPI) verilerinden elde edilen veri setleri kullanılmıştır. Her düğüm için komşu sayısı, Adamic Adar endeksi, Jaccard katsayısı ve etiket değerleri de dahil olmak üzere yerel indeks değerlerini hesaplayarak türetilen veriler oluşturulmuştur. Başarı oranlarını artırmak için, bir model, tüm veri setlerinde ön işleme tekniği olarak kesikli yöntemi uygulayan bir model geliştirilmiştir. Bayesian Ağları, Çok Katmanlı Algılayıcı (MLP), Rastgele Orman ve k-En Yakın Komşuluk (kNN) gibi makine öğrenimi algoritmaları karşılaştırılmış ve değerlendirilmiştir. Sonuçlar, en yüksek kenar sayısına sahip olan Twitch veri setinin başarılı sonuçlar verdiğini göstermektedir. Ağ yapısındaki kenar sayısının performansa katkısı vurgulanmış ve kesikli yöntemin uygulandığı veriler için daha başarılı metrik değerler elde edildiği gözlemlenmiştir.

Anahtar kelimeler: Bağlantı Tahmini, Protein-Protein Etkileşimi, Sosyal Ağlar, Makine Öğrenme Algoritmaları, Biyolojik Ağlar.

1. INTRODUCTION

Complex network analysis is becoming more and more popular these days due to a number of factors, including the incidence of epidemic diseases, technology breakthroughs in healthcare, and the retail industry's comeback as a result of population expansion [1]. In a variety of fields, such as healthcare, social media, and retail, network-based systems are essential, especially in the areas of medication creation, recommendation systems, and treatment approaches [2-3]. Link prediction

*Sorumlu Yazar (Corresponding Author)
e-posta : hasibe.candan@btu.edu.tr

becomes more important in this setting, especially when it comes to tasks like determining the likelihood of connections between two nodes and discovering missing links. Link prediction, one of the most talked-about subjects in complex network analysis, provides answers to a number of problems [4]. Finding a network's missing links has wide-ranging effects. Predicted links, for example, can improve user experience and increase sales in product recommendation systems [5]. Link prediction in social media makes it easier to make new connections or get in touch with old friends [6]. Predicting connections within human cells is another area of study. By forecasting protein interactions between cells, these studies seek to protect cells from possible microscopic dangers that could cause abnormalities [7]. Various link prediction methods exist for estimating protein-protein interactions (PPI), determining the similarity ratio between two proteins, estimating nodes within graph structures, and measuring the probability of paths between two nodes [8]. The evolution of protein interactions within a protein complex involved in reactions was the subject of another study that examined biological functions. To evaluate signal dynamics in protein interaction networks, a specific mass spectrometry-based reaction monitoring technique was developed. The goal was to use interactions to find important core proteins and new interconnected networks [9-10]. Scoring missing links in a network using certain techniques, such as mathematical formulas and methodologies like the Adamic-Adar index, Jaccard index, and network topology, is a popular practice in this field of study [11–12]. Instead of depending only on ranking techniques, some studies tackle this by building a machine learning model [13]. Furthermore, contemporary research uses graph embedding methods, supervised learning, and artificial intelligence algorithms [12-14]. Approaches in this field are categorized into in vivo and in vitro. In computational contexts, link prediction depends on topological structures, just like in lab-based research. Anomaly detection, which uses link prediction algorithms to find malevolent users [15], is the focus of some studies, while others investigate transfer link prediction in heterogeneous networks to create recommendation systems [16]. Recent approaches that tackled issues including multidimensional networks, scalability, and network dynamics were reviewed in an article. Among graph-based metrics, Common Neighbors [18], Adamic-Adar [19], Diffisuion Alignment Coefficient[17] and Katz [20] are often employed measures. A citation network dataset was also used to suggest a probabilistic method that included attributes including author names, summaries, and locations. Additionally, the Adamic-Adar index, Jaccard coefficient, label value, and number of neighbors were included in a new dataset [21]. The study sheds light on several connection analysis model approaches. Digital databases from in vitro settings are used in in vivo investigations, which use a variety of topological components to create new metrics that are appropriate for the requirements of the applications. It has been underlined that link prediction techniques may not work as well in some network types but do well in others. Linkage analysis of graph topologies produced by SVM on gene-protein-protein interactions and social network datasets is one area of research in this field. Furthermore, graph-based metrics and local indices were computed, producing a reconstructed dataset [21]. The newly created dataset, along with connection prediction datasets and graph-based metrics, was input into a machine learning model and tested separately. One of the most common estimation algorithms based on mathematical models is the Kalman filter [23-24]. Different artificial intelligence approaches have been proposed in the literature for complex forecasting problems [22]. However, implementing these algorithms can be challenging due to computational complexity, making machine learning techniques more advantageous [26]. Classifier algorithms such as Bayesian Net, Multilayer Perceptron, k-Nearest Neighborhood, and Random Forest were employed. The impact of discretization as a pre-processing method on model performance was examined, and accuracy, precision, recall, and AUC values were compared across datasets. The results highlight the most effective method in this study.

The majority of studies in the literature have taken a unilateral approach, focusing solely on local, global, or algorithmic methods. This study, however, employs machine learning techniques to construct similarity in link prediction algorithmically. Additional insights are also provided by datasets that are generated using values obtained from local indices. Local index values are used to develop attributes in these derived datasets. Both the network data and the newly created datasets from it are subjected to a variety of classifier techniques. Key metric values are used for evaluation in order to provide a thorough interpretation. This study adds to the literature by effectively exposing performance results for both protein-protein interaction network data and social network data from a fresh perspective, going beyond the conventional one-way link prediction methodologies that focus on global and local indices.

2. MATERIAL AND METHODS

2.1. Model Structure

Derived data or network data was utilized in the created model structure, as detailed in the dataset section. The discretization method was applied as a preprocessing step, and the study evaluated the impact of discretization on success. To assess this impact, metric values were compared with and without discretization applied. Subsequently, machine learning algorithms were employed and evaluated using the metrics illustrated in Figure 1.

2.2. Link Prediction

Link prediction involves forecasting the connection between two nodes within a network. Examples of link prediction include predicting friendship status among

users in a social network or estimating interactions between genes and proteins in a biological network [25].

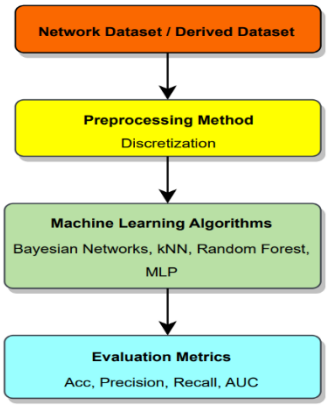

Figure 1. Created model structure

Illustrated in Figure 2, link prediction is the process of foretelling future links between unlinked pairs based on existing connections. For instance, if Node 1 is connected to both Node 2 and Node 3, and Nodes 2 and 3 are not directly linked, link prediction aims to forecast a connection between Nodes 2 and 3, considering Node 1 as a common node. Link forecasting not only analyzes future links but also predicts missing node connections within a given time period. The probability of a connection between nodes increases as their similarity rises, with measures like cosine similarity and Euclidean distance serving as examples [26-27]. Maintaining a similar number of nodes and edges is crucial, ensuring that datasets accurately represent networks of the same type. This study compared two distinct link prediction approaches: the machine learning approach, employing algorithms, and the local and global index approach, employing similarity measures [28].

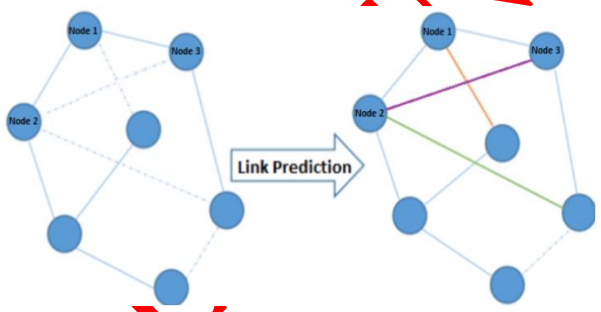

Figure 2. Link prediction

In Table 1 of this study, the quantities of nodes and edges within the datasets are provided. The quantity of nodes and edges within the utilized datasets holds significance. Examining the variations in edge and node counts offers crucial insights to enable commentary on results, particularly in networks that share similarities but differ only in terms of node-edge numbers. We elaborated on this aspect in the Experimental Results section.

Local indices are one of the simplest approaches applied to calculate the similarity score, taking into account the number of neighboring nodes and the degree of neighborhood for link prediction.

Table 1. Number of nodes and edges of data sets

	Nodes	Edges
Facebook Net	4039	88234
Human PPI	21557	342353
Twitch Net	34118	429113
Twitter Net	38918	32523

When the path distance is less than two, a node is generally considered a neighboring node. Some of the local indices are measured with values such as the Jaccard index, Adamic Adar index, and Resource Allocation index. Local similarity indices are widely used in real applications as they maintain optimal prediction performance while reducing resource usage and computational complexity. Global indices, on the other hand, calculate the similarity score based on the global connectivity structure of the graph where the path distance between nodes is more than two. Global indices utilize the entire topological information network to score each connectionGlobal indices find all direct and indirect paths that are interesting to include in the similarity score, in contrast to local index approaches. The great dimensionality of the networks makes global similarity indexes expensive and time-consuming when used for link prediction in big networks. In this study, machine learning algorithms from algorithmic approaches, along with global and local approaches for link prediction, were employed. The values in local approaches are involved in creating the derived dataset.

2.2.1. Adamic adar index

The Adamic-Adar index is a widely used metric for link estimation and aims to capture the strength of relationships between nodes that share common neighbors. This metric recognizes that common neighbors with fewer links (low degree) are more important in estimating the links between two nodes. Using a logarithmic function, the Adamic-Adar index reduces the influence of neighbors with a large number of connections, making the metric more sensitive to the presence of less connected common neighbors. The value of this index increases as the number of common neighbors between two nodes increases, but the influence of each common neighbor is adjusted by their degree. The rationale behind this is that less connected neighbors are more likely to provide more valuable information for predicting new connections between nodes, compared to more connected neighbors that already have well-established relationships in the network [29]. This approach is particularly effective in sparse networks, as it is more effective for identifying less connected but important relationships in the network. Furthermore, the Adamic-Adar index has applications in diverse fields such as social networks and biological networks, where it reveals hidden connections and interactions that are not immediately apparent by direct observations, but help to predict potential interactions between nodes. By filtering

out the influence of high-ranking nodes, this metric emphasizes rare and less obvious connections, making it a powerful tool for predicting future connections in complex networks [30].

$$s(u, v) = \sum i \in Nu \cap Nv \; 1/\log 2 \; (|Ni|) \qquad (1)$$

Formula 1 calculates an aggregate where each common neighbor is weighted by the inverse of the logarithm of its degree. Low-degree neighbors receive higher weights, while the weights of high-degree neighbors decrease. This is based on the assumption that low-degree neighbors are more important in link prediction. In terms of its relation to network topology, the Adamic-Adar index estimates connectivity probabilities taking into account the local structure of the network. In particular, the distribution and degrees of common neighbors provide important clues about the information flow and interactions in the network. Therefore, the Adamic-Adar index is an effective method to detect hidden or weak links in the network.

2.2.2. Jaccard coefficient

Jaccard coefficient is a similaritymeasure obtained by dividing the number of common features by the total number of features in two sets after feature extraction. It was developed to compare two sets and represents the ratio of the number of common neighbors of two nodes to the total number of neighbors [31]. The Jaccard coefficient is particularly effective for link prediction because it captures the intuition that nodes sharing many common neighbors are more likely to form connections themselves (reflecting network transitivity); it normalizes by the total number of neighbors, making it robust against degree heterogeneity in networks; it effectively utilizes the network's structural information by focusing on neighborhood overlap; it offers computational efficiency through simple set operations; and it demonstrates versatility across diverse network types including social, biological, and information networks, making it a reliable feature for link prediction across domains. The Jaccard coefficient is calculated by taking the intersection of the neighbor sets N(u) and N(v) and dividing it by their union. It is shown in Formula 2.

$$s(u, v) = |Nu \cap Nv| \; / \; |Nu \cup Nv| \qquad (2)$$

2.3. Classifier Methods

2.3.1. Support vector machine

Support Vector Machines (SVM) is explained by defining the formation of a hyperplane to optimally separate two classes. For classification, it relies on the concept of delineating two clusters on a plane by drawing a boundary. The SVM algorithm decides which data points belong to which class boundary [32]. SVM is a supervised classification algorithm based on statistical learning theory. Initially designed for the classification of two-class linear data, the mathematical algorithms of SVM were later extended to handle multi-class and non-linear data classification. The working principle of SVM involves estimating the most appropriate decision function to distinguish between two classes and defining the hyperplane that best separates the two classes [36].

2.3.2. Bayesian networks

Random is a graph-based probabilistic analysis method that defines the statistical relationship between variables. Probabilistic analysis methods are employed to determine if the current observation aligns with one of its hypotheses [33]. G = (V, E) represents a directed, noncyclic graph mapped to random variables V = {V1, V2, … , Vn } where vertices denote the conditional probabilities between E and directed edges E. The Naive Bayes classifier is a nonlinear, Bayesian theorem-based probabilistic classification algorithm capable of processing complex data and providing the probability of a new event occurring based on an event [34]. This algorithm assumes that a feature belonging to a class in the dataset is independent of other features, which may lead to suboptimal performance. Despite its simplicity and speed compared to other algorithms, the Naive Bayes algorithm is fast, easy to implement, and more resistant to overfitting, as it typically estimates fewer parameters than other classifiers [35].

2.3.3. k- nearest neighbor

This involves multi-label learning, where the class of the relevant instance is determined based on the class of the k nearest neighbors (kNN), consisting of associated instances. The method operates by analyzing training samples with given sets of tags and predicting tag sets for samples whose class is undetermined. It utilizes the Euclidean distance for calculation. kNN stands as one of the oldest and simplest pattern recognition methods [39]. It classifies each unlabeled sample by considering the predominant label among its k-nearest neighbors in the training set. Its execution relies heavily on the distance metric.

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} wr \; (ar)(xi) - ar(xj))2} \qquad (3)$$

2.3.4. Random forest

The Random Forest classifier method is a ensemble learning algorithm that constructs numerous decision trees for regression and classification tasks, making class predictions based on the nature of the problem [37]. This methodology operates by generating multiple decision trees and amalgamating them to produce a more accurate and stable forecast. It is utilized for both classification and regression problems within machine learning methods [39]. The Random Forest algorithm allows for the creation of diverse models, generating classifications by training each decision tree on a different subset of observations across multiple trees. Widely applied in both classification and regression scenarios, a key feature of the algorithm is its ability to gain insights into datasets by constructing various models on the same dataset.

2.3.5. Multilayer perceptron

The Multilayer Perceptron (MLP) comprises artificial neurons connected to each other, resembling neurons in the brain. Information flow is facilitated through the connections of these artificial neurons. The MLP structure includes input, intermediate, and output layers. Initially, inputs are acquired and transmitted to the intermediate layers, which can be one or more. The outputs are determined as a result of the activation function [41].

2.4. Discretization

Discretization is a preprocessing method used to convert continuous values into discrete categories. This technique makes data sets denser and more condensed, which gives machine learning models a better structure and simplifies data analysis by narrowing the range of data. Four steps typically make up the discretization process: sorting continuous values, determining a cut-off point among them, separating or merging continuous intervals based on these cut-off points, and determining the points at which the splitting process will conclude [40]. Through the reduction of noise caused by continuous characteristics, this procedure makes the data easier to handle and enhances the performance of link prediction models. Compared to other pre-processing methods such as normalization or standardization, discretization provides a more robust analysis by breaking the data into more structured and meaningful categories. This often improves the accuracy of the model and leads to more reliable results.

3. FINDINGS

3.1. Confusion Matrix and Metrics

To evaluate the performance of classification models in machine learning, the confusion matrix, which compares actual and predicted values, is often employed [43]. It distinguishes between correct predictions (true positives and true negatives) and errors (false positives and false negatives). Experimental results were acquired using the 10-fold cross-validation test technique [44-45]. In this technique, the dataset is divided into 10 equal folds, nine for training and one for testing in each iteration, resulting in a total of 10 tests. The performance values gathered from these tests are then averaged [42]. Confusion matrix provides a comprehensive indicator for evaluating the performance of a classification model, analyzing errors, and improving the model [46]. Confusion matrix is given in Figure 3.

		Prediction	
		P	N
Real	P	TP	FN
	N	FP	TN

Figure 3. Confusion matrix

Class accuracy provides the proportion of correctly predicted values, as expressed in Equation 4. The precision value indicates the stability of predictions, representing the proportion of correct predictions for the class (TP) to be predicted as the target, and is given by Equation 5. Recall is a performance metric calculated by dividing the number of true positives by the sum of true positives and false negatives, as shown in Equation 6. It measures the effectiveness of the model in recognizing a class, and a high Recall value, as depicted in Equation 7, indicates that the model is less likely to miss instances of that class. The AUC value corresponds to the ROC curve plot and ranges from 0 to 1; the closer it is to 1, the better the model's performance.

$$\text{Accuracy} = (TP + TN)/(P+N) \quad (4)$$

$$\text{Precision} = TP/(TP+YN) * (TP+FP) \quad (5)$$

$$\text{Recall} = TP/(TP + FN) \quad (6)$$

$$\text{AUC} = 1/2 * (TP/(TP+ FN) + TN/(TN+ FP)) \quad (7)$$

3.2. Data Sets

The label values for a dataset comprising four network structures were determined using Support Vector Machines. The initial dataset consists of three attribute values: the first node, the second node, and the label.

In the subsequent derived datasets, new values are obtained from the data, including Jaccard coefficient value, neighborhood value, Adamic Adar index value, resource allocation index, and label. A new dataset was generated by acquiring values from local indices. This process results in the creation of a new dataset with these attributes. The datasets used include the Facebook dataset, Twitch dataset, Human Protein-Protein Interaction (PPI) dataset, and Twitter dataset, all obtained from http://snap.stanford.edu/data/[47].

The data sample numbers are given in Table 2. Besides the number of samples, the number of nodes and edges in the data sets is important and is given in Table 1 of the Link Prediction section.

Table 2. Data sets

Data Sets	Number of Samples
Facebook Network	1410
Human PPI	566
Twitch Network	6158
Twitter Network	7710

3.3. Experimental Results

Table 3 illustrates the performance metrics for link estimation on the non-disaggregated data, while Table 4 provides performance values for the data after applying discretization. The initial format includes data from the

Facebook Network, Human PPI, Twitch Network, and Twitter Network. In the second format, referred to as derived data, new values such as Jaccard coefficient, Neighborhood value, Adamic Adar index, Resource Allocation index, and label are added one by one to the data derived from these datasets, which originally includes the first node, the second node, and the label value indicating the connection status between these two nodes. A new dataset is formed through these calculations.

Derived data appears to yield better results across all datasets except for the Twitch dataset. Consequently, despite not achieving exceptionally high performances, the derived data proves effective in enhancing the overall performance rate. In the Twitch dataset, the Random Forest algorithm delivered the most successful classifier result in the link prediction method without discretization, achieving a class accuracy of 92.14% on the network dataset. The Bayesian Networks algorithm also produced one of the top results in the Twitch dataset, with a rate of 88.91%. Overall, it can be inferred that the Twitch dataset, with 34,118 nodes and 429,113 edges as indicated in Table 1, exhibits a higher level of connectivity compared to the other datasets. It is evident that Bayes-based algorithms consistently provide successful outcomes. Algorithms rooted in probability theory have been observed to be more successful irrespective of network type and edge-node numbers, while neighborhood-based algorithms exhibit a lower success rate. It can be concluded that algorithms based on probability theory, such as Bayesian Networks, consistently yield better results in such highly connected networks. On the other hand, neighborhood-based algorithms like kNN and MLP show comparatively lower success rates. This indicates that the network's structural properties, including its size and connectivity, significantly influence the model's performance.

Table 4 shows the link prediction performance metrics on the disaggregated dataset. All performance ratios increased on the discretized dataset. In this context. it is clear that the performance of the model improves with the discretized data compared to the non-discretized data. This shows the effectiveness of using discretization in the model. In Bayesian Networks, while the performance value was 49% when no discretization was applied to Facebook data, this value increased to 72% with discretization. Comparing the values in Table 3 and Table 4, it can be seen that the derived datasets perform better than the applied datasets when no discretization preprocessing is applied. Both the derived dataset and the discretization network perform well on all datasets. The significant improvement in performance with discretized datasets arises from the more distinct representation of relationships between nodes, which allows for a more effective learning of the underlying structure. Considering all datasets with discretization applied, there is an increase from 88% to 90.7% on the Twitch dataset, which gives the best results. Similarly, MLP and kNN also gave the best results on the same dataset.

Table 3. Linkage estimation on the non-disaggregated data set

Data Set	Bayesian Networks				kNN			
	Acc(%)	**Precision**	**Recall**	**AUC**	**Acc(%)**	**Precision**	**Recall**	**AUC**
Facebook Net	49.47	0.496	0.693	0.493	55.3004	0.553	0.558	0.553
Facebook Derived	64.72	0.789	0.936	0.724	42.41	0.403	0.601	0.4
Human PPI Net	49.47	0.491	0.693	0.493	55.3	0.554	0.548	0.553
Human Derived	65.48	0.74	0.883	0.696	40.28	0.383	0.403	0.393
Twitch Net	**88.91**	**0.965**	**0.889**	**0.919**	66.33	0.649	0.71	0.662
Twitch Derived	66.67	0.688	0.614	0.748	42.03	0.385	0.42	0.482
Twitter Net	52.64	0.533	0.526	0.542	51.95	0.519	0.523	0.519
Twitter Derived	65.82	0.799	0.949	0.725	41.9	0.395	0.419	0.427
	Random Forest				**MLP**			
Facebook Net	59.364	0.596	0.608	0.626	50	0.506	0.707	0.5
Facebook Derived	39.1017	0.387	0.55	0.492	65.86	0.778	0.682	0.736
Human PPI Net	59.36	0.591	0.608	0.626	50.35	0.506	0.504	0.5
Human Derived	40.363	0.4	0.4	0.501	65.6	0.665	0.656	0.707
Twitch Net	**92.14**	**0.939**	**0.901**	**0.947**	50.5196	0.507	0.505	0.51
Twitch Derived	42.42	0.414	0.424	0.523	66.66	-	0.667	0.594
Twitter Net	54.46	0.545	0.542	0.564	55.03	0.55	0.558	0.577
Twitter Derived	41.8	0.416	0.418	0.502	66.831	0.655	0.412	0.748

Table 4. Connection estimation on disaggregated data set

Discretized Data Set	Bayesian Networks				kNN			
	Acc(%)	**Precision**	**Recall**	**AUC**	**Acc(%)**	**Precision**	**Recall**	**AUC**
Facebook Net	**72.1986**	**0.743**	**0.678**	**0.699**	**76.45**	**0.799**	**0.596**	**0.793**
Facebook Derived	65.3428	0.8	0.952	0.729	67.32	0.533	0.159	0.723
Human PPI Net	49.47	0.491	0.297	0.493	49.47	0.494	0.297	0.493
Human Derived	66.077	0.743	0.823	0.721	64.66	0.604	0.216	0.705
Twitch Net	**90.711**	**0.92**	**0.82**	**0.931**	**90.71**	**0.92**	**0.82**	**0.932**
Twitch Derived	66.6775	0.687	0.667	0.749	66.7316	0.697	0.667	0.747
Twitter Net	53.25	0.534	0.435	0.548	54.74	0.642	0.139	0.548
Twitter Derived	66	0.804	0.956	0.727	67.8	0.697	0.678	0.749
	Random Forest				**MLP**			
Facebook Net	**76**	**0.799**	**0.596**	**0.793**	**74.2553**	**0.747**	**0.677**	**0.799**
Facebook Derived	67.32	0.533	0.159	0.724	65.6738	0.488	0.631	0.737
Human PPI Net	49.6466	0.496	0.399	0.496	50	0.5	0.3	0.499
Human Derived	64.42	0.614	0.283	0.705	65.84	0.663	0.512	0.725
Twitch Net	**90.711**	**0.92**	**0.82**	**0.931**	**90.69**	**0.919**	**0.823**	**0.923**
Twitch Derived	66.6	0.705	0.69	0.748	66.94	0.681	0.567	0.753
Twitter Net	64.7341	0.758	0.139	0.548	54.06	0.555	0.287	0.557
Twitter Derived	67.93	0.702	0.679	0.75	67.0558	0.668	0.492	0.75

The higher performance of the Twitch dataset is attributed to its higher connectivity, which provides more detailed relationships between nodes, thereby enhancing the model's predictive capabilities. It can be concluded that discretization works effectively with MLP and kNN. Facebook data emerged as the second best performing dataset. The difference in performance between the Twitch and Facebook datasets underscores how network topologies, such as the level of connectivity between nodes, influence the performance of link prediction models. Among all datasets, Twitch exhibited the best performance concerning the number of edges and nodes in both Discretized and Non-Discretized datasets. In general, it can be stated that as the number of edges decreases compared to the number of nodes, the performance decreases, irrespective of the network's content. This suggests that models perform better on highly connected networks, where the relationships between nodes are more easily captured.

4. CONCLUSION

Recently, comprehensive network analysis has gained popularity across various domains such as education, social platforms, shopping, treatment methods, and drug development. Link prediction plays a crucial role in identifying missing connections within networks, offering multiple outcomes. Different link prediction methods, based on various dimensions, include predictive bonds in protein-protein interaction (PPI), connections between two proteins, prediction of nodes within graph structures, and estimation of paths between two nodes.

In this study, Support Vector Machines were utilized for network analysis in predicting connectivity within protein-protein networks and social structures, using datasets from Twitch, Facebook, Twitter, and Human-Gene PPI. Linkage analysis involved classifier reflections on the dataset with the link map, calculating the probability of linkage. Derived data, obtained through similarity-based methods, included calculations of the number of neighbors for each node, Adamic Adar index, Jaccard parameters, and label values. Analysis was conducted on the obtained dataset using classifier machine learning methods. To assess performance, the preprocessing method, Discretization method, was applied to the generated dataset. In two models, one with and one without preprocessing, each of the network datasets underwent derivation and was subjected to classifier algorithms using only the network dataset. The evaluation was carried out using classifier metrics, with Bayes Net, Multilayer Perceptron, k-Nearest Neighborhood, and Random Forest employed as machine learning algorithms. The model showed its success at the highest level in the Twitch data set. The highest performances were consistently achieved across all methods and datasets when Discretization was applied. The study, explored the impact of the number of nodes and edges on performance rates, revealing that the Twitch dataset with the highest number of edge nodes demonstrated superior performance. It was observed that

discretization enhances success rates in finding and producing data in social networks.

In future work, there is a plan to offer a broader perspective by considering different classifiers on newly derived data sets created by calculating global index values.

AUTHORS' CONTRIBUTIONS

Hasibe CANDAN KADEM: Wrote the code, carried out the experiments, contributed to the study's design, and produced the manuscript.

Volkan ALTUNTAŞ: Contributed to the idea, study design, and overall supervision, reviewed and revised the manuscript.

DECLARATION OF ETHICAL STANDARDS

There is no need to obtain the prepared medical ethics committee permission. There is no conflict of interest with any person/institution in the article prepared.

CONFLICT OF INTEREST

There is no conflict of interest in this study

REFERENCES

[1] Altuntas, V., Gok, M., & Kocal, O. H. "Response of Lyapunov exponents to diffusion state of biological networks". ***International Journal of Applied Mathematics and Computer Science*** 30(4), 689-702, (2020).

[2] Orman GK. "Discovering Link Prediction Methods' Performances by Network Topology Relation". ***Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi***, 22(4), 778-788, (2022).

[3] Kösesoy İ, Gök M, Kahveci T. "Prediction of Host-Pathogen Protein Interactions by Extended Network Model". ***Turkish Journal of Biology***, 45(2), 138-148, (2021).

[4] Lei C, Ruan J. "A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity". ***Bioinformatics***, 29(3), 355-364, (2013).

[5] Kaya B. "Hotel recommendation system by bipartite networks and link prediction". ***Journal of Information Science***, 46(1), 53-63, (2020).

[6] Shabaz M, Garg U. "Predicting future diseases based on existing health status using link prediction". ***World Journal of Engineering***, (2021).

[7] Zareie A. "Sakellariou R. Similarity-based link prediction in social networks using latent relationships between the users". ***Scientific Reports***, 10(1), 1-11, (2020).

[8] Bandyopadhyay S, Chiang, CY. Srivastava J, Gersten M, "White S, Bell R, Ideker T, A human MAP kinase interactome". ***Nature Methods***, 7(10), 801-805, (2010).

[9] Kösesoy, İ, Gök M, Öz C. "A new sequence based encoding for prediction of host–pathogen protein interactions". ***Computational Biology and Chemistry***, 78, 170-177, (2019).

[10] Bisson, N, James, D. A, Ivosev G, Tate S. A, Bonner R, Taylor L, Pawson T. "Selected reaction monitoring mass spectrometry reveals the dynamics of signaling through the GRB2 adaptor". ***Nature Biotechnology***, 29(7), 653-658, (2011).

[11] Martínez V, Berzal F, Cubero JC. "A survey of link prediction in complex networks". ***ACM Computing Surveys (CSUR)***, 49(4), 1-33, (2016).

[12] Karaahmetoğlu, E, Ersöz, S, Türker, A. K., Ateş, V., İnal A. F. "Evaluation of Profession Predictions for Today and the Future with Machine Learning Methods: Emperical Evidence From Turkey". ***Politeknik Dergisi***, 26(1), 107-124, 10.2339, (2023).

[13] Altuntas, V. "NodeVector: A Novel Network Node Vectorization with Graph Analysis and Deep Learning". ***Applied Sciences***, 14(2), 775, (2024).

[14] Yücel, M., Osmanca, M. S. and Mercimek, İ. F. "Machine learning algorithm estimation and comparison of live network values of the inputs which have the most effect on the FEC parameter in DWDM systems". ***Politeknik Dergisi***, 1-1, (2024).

[15] Calp, M. H., & Bütüner, R. Detecting the Cyber Attacks on IoT-Based Network Devices Using Machine Learning Algorithms. ***Politeknik Dergisi***, *1-1*, 1340515, (2024).

[16] Wang, M, Qiu L, Wang X. "A survey on knowledge graph embeddings for link prediction". ***Symmetry***, 13(3), 485, (2021).

[17] Baskar, P, Joseph, MA, Narayanan, N, Loya, RB. "Experimental investigation of oxygen enrichment on performance of twin cylinder diesel engine with variation of injection pressure". ***In 2013 International Conference on Energy Efficient Technologies for Sustainability (pp. 682-687) IEEE.*** Nagercoil, India, (2013).

[18] Altuntas, V. "Diffusion Alignment Coefficient (DAC): A Novel Similarity Metric for Protein-Protein Interaction Network". ***IEEE/ACM Transactions on Computational Biology and Bioinformatics***, 20(2), 894-903, (2022).

[19] Zeng S. "Link prediction based on local information considering preferential attachment". ***Physica A: Statistical Mechanics and its Applications,*** 443, 537-542, (2016).

[20] Deylami H. A. "Asadpour M. Link prediction in social networks using hierarchical community detection". ***In 2015 7th Conference on Information and Knowledge Technology (IKT) (pp. 15). IEEE***, Urmia, Iran, 26-28 May (2015).

[21] Rattigan M. J., Jensen D. "The case for anomalous link discovery". **ACM Sigkdd Explorations Newsletter**, 7(2), 41-47, (2005).

[22] Qian F, Gao Y, Zhao S, Tang J, Zhang Y. "Combining topological properties and strong ties for link prediction". ***Tsinghua Science and Technology***, 22(6), 595-608, (2017).

[23] Kadem, O., Candan, H., and Kim, J. "Hybrid Deep Neural Network for Electric Vehicle State of Charge Estimation". ***IEEE 3rd International Conference on Electrical Power and Energy Systems (ICEPES) (pp. 1-6)***, (2024).

[24] Kadem O, Kim J. "Mitigation of state of charge estimation error due to noisy current input measurement". ***Proceedings of the Institution of Mechanical Engineers, Part I Journal of Systems and Control Engineering***, (2023).

[25] Kovács I.A., Luck K, Spirohn, K, Wang Y, Pollis, C, Schlabach S, Barabási A. L. "Network-based prediction of protein interactions". ***Nature Communications***, 10, 1-8, (2019).

[26] Kadem O. "Real-Time State of Charge Estimation Algorithm for Electrical Batteries". ***PhD thesis, University of Leeds.*** https://etheses.whiterose.ac.uk/id/eprint/31973, (2022).

[27] Çakmak E, Kaya B, Kaya M. "İki Parçalı Ağda Bağlantı Tahminine Dayalı İlgi Çekici Nokta Tavsiyesi". ***Computer Science, (Special)***, 154-161, (2021).
[28] Zhang M.L., Zhou Z.H. "ML-KNN: A lazy learning approach to multi-label learning". ***Pattern Recognition***, 40(7), 2038-2048, (2007).
[29] Martínez B, Cubero, Martínez V, Berzal F, Cubero JC. "A survey of link prediction in complex networks", ***ACM Computing Surveys (CSUR)***, 49(4), (2017).
[30] Adamic, LA, & Adar E. "Friends and neighbors on the web". ***Social Networks***, 25, 211-230, (2023).
[31] Jaccard P. "The distribution of the flora in the alpine zone". **1.** ***New Phytologist***, 11(2), 37-50, (1912).
[32] Clauset A, Moore C, Newman M.E. "Hierarchical structure and the prediction of missing links in networks". ***Nature***, 453(7191), 98-101, (2008).
[33] Jiang S, Xu K and Xiao J," Link Prediction by Combining Local Structure Similarity with Node Behavior Synchronization", ***IEEE Transactions on Computational Social Systems***, vol. 11, no. 3, pp. 3816-3825, (2024).
[34] Liu J, Li B and Dillon T, "An improved naive Bayesian classifier technique coupled with a novel input solution method [rainfall prediction]", ***IEEE Transactions on Systems, Man, and Cybernetics, Part C*** (Applications and Reviews), vol. 31, no. 2, pp. 249-256, (2001).
[35] Baras D, Fine, S, Fournier L et al. "Automatic boosting of cross-product coverage using Bayesian networks". ***Int J Softw Tools Technol Transfer 13***, 247–261., (2011).
[36] Mathur A, Foody G.M. "Multiclass and binary SVM classification: Implications for training and classification users". ***IEEE Geoscience and Remote Sensing Letters*** (5):241–245, (2008).
[37] Oshiro TM, Perez PS, Baranauskas JA. "How many trees in a Rotation Forest". ***In International Workshop on Machine Learning and Data Mining in Pattern Recognition*** (pp. 154-168), *Springer*, Berlin, Heidelberg, Germany, (2012).
[38] Vishwakarma, M., Kesswani, N. "A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection". ***Decision Analytics Journal***, *7*, 100233, (2023).
[39] Pal M. "Random forest classifier for remote sensing classification". ***International Journal of Remote Sensing***, 26(1), 217-222, (2005).
[40] Fayyad U, Irani K. "Multi-interval discretization of continuous-valued attributes for classification learning", ***International Joint Conference on Artificial Intelligence,*** (1993).
[41] Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. "Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions". ***Computer Science Review,*** *38*, 100311, (2020).
[42] LeCun Y, Jackel L, Bottou L, Brunot A, Cortes C, Denker J, Vapnik V. "Comparison of learning algorithms for handwritten digit recognition". ***In International conference on artificial neural networks*** (Vol. 60, No. 1, pp. 53-60), Australia, (1995).
[43] Sharma V, Yadav S, Gupta M, "Heart Disease Prediction using Machine Learning Techniques". ***2nd Int. Conf. Adv. Comput. Commun. Control Networking, IEEE, ICACCCCN***, Greater Noida, India, (2020).
[44] Katarya R, Meena S.K.," Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis". ***Health Technol***. vol. 11, no. 1, pp. 87–97, 10.1007/s12553-020-00505-7, (2021).
[45] Mijwil M.M., Abttan R.A. "Utilizing the genetic algorithm to pruning the C4. 5 decision tree algorithm". ***Asian Journal of Applied Sciences***, 9(1), (2021).
[46] Valero-Carreras, D., Alcaraz, J., & Landete, M., "Comparing two SVM models through different metrics based on the confusion matrix". ***Computers & Operations Research***, *152*, 106131, (2023).
[47] Leskovec, J., Sosič, R. "Snap: A general-purpose network analysis and graph-mining library". ***ACM Transactions on Intelligent Systems and Technology (TIST)***, *8*(1), 1-20, (2016).