International Journal of Analysis and Applications
Volume 17, Number 4 (2019), 578-585
URL: https://doi.org/10.28924/2291-8639
DOI: 10.28924/2291-8639-17-2019-578

IMPROVED LINK PREDICTION USING PCA

ANKITA[1,*], NANHAY SINGH[2]

[1] *University School of Information Communication and Technology, Guru Gobind Singh Indraprastha University, New Delhi, India*

[2] *Ambedkar Institute of Advanced Communication Technologies and Research, Guru Gobind Singh Indraprastha University, New Delhi, India*

[*] *Corresponding author:* ankita.it.13@gmail.com

Abstract. Link Prediction is known as a challenging problem in the area of online social media. Earlier, learning model for link prediction task has been proposed by many researchers. But the classification of imbalanced and high dimensional data is an interesting and challenging problem in machine learning due to presence of unbalanced and redundant or correlated data which break down the classification performance. In this paper, we have balanced the data and used Principle Component Analysis (PCA) to reduce the correlated data and improved the performance of link prediction model. Experiment is carried out on social network data set and the use of PCA method has improved the performance in classification of links.

1. Introduction

Link Prediction Problem: Online social media is a structure of users, where nodes are the users or entity and edges represent collaboration, interaction or association between users. Link prediction is known as a fundamental problem in the online social network where task is to predict links in the near future. It can be defined as a given social network graph consisting of nodes as users and links between them at time t, predict new links between users at time t'(where t is less than t') as shown in Figure 1. As new relationship

Received 2019-03-04; accepted 2019-04-05; published 2019-07-01.

2010 *Mathematics Subject Classification.* 91D30.

Key words and phrases. feature; reduction; link; learning; prediction .

keeps on adding in social network and old relations may be deleted, this is considered as the challenging problem.

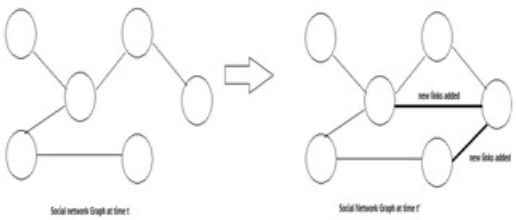

FIGURE 1. Link Preditcion example

There are many different approaches proposed by researchers in the past and applied in different areas. For a co-authorship network, using link prediction, future association between authors can be predicting. Similarly in online social network, the friend connections between the users can be identified or predict. In the area of spam mail detection, link prediction model can be used to detect anomalies in the emails [1]. The link prediction techniques also have been applied in disease prediction by Folino and Pizzuti [2]. They have applied link prediction to predict the diseases. The diseases are represented as the nodes and the incidence of the diseases in a patient is represented by an edge. Using link prediction, a score between two diseases is identified and using link score, probability of co-occurrence of the diseases is identified. Link prediction model has also been applied in recommender system [3,4], where new product can be recommended to users based on user preferences or rating. Due to wide range of application in different areas, link prediction problem has attracted more attention for the researchers. The link prediction methods can be classified as similarity based methods or learning based methods. In the case of similarity based model, similarity score is assigned on the edge of the network by using similarity metrics and used to predict links between two nodes. In [5] [6], similarity based model has been used by authors to predicting links in the graph network. In learning based model the link prediction is considered as the classification problem and in [7], supervised learning model used to predict links in the social network graph. Most of the existing work have used the static nature of the network and have used the supervised learning algorithm using topological features of the network without using any dimension reduction techniques to improve the feature set. In this paper, we propose a supervised learning algorithm where features vector created using network information and to improve the performance of learning based model we have balance the data and used Principle component analysis (PCA) for dimension reduction. Principle Component Analysis is a method to convert the attributes of a dataset into uncorrelated datasets. The components of this new dataset are known as principle components (PCs). PCA is used to reduce the high dimensional data into low dimension datasets having high variance. Highly correlated attributes in a dataset can degrade the performance in prediction model.

The contribution of this paper: 1) We proposed a learning based model using features based model. 2) We have balanced the data using oversampling and used dimension reduction technique, PCA to improve the prediction accuracy. 3) We have used karate club dataset in our experiment to show effectiveness of our approach and used Accuracy, Recall, Precision and F-measure for performance evaluation. Here, paper is ordered as: section 2 – the related work; section 3 – proposed methodology for solving link prediction problem; section 4 – experimental results and discussion; section 5 – conclusion of the work.

2. Related Work

Liben-Nowell and Kleinberg [8] have described Link prediction problem, where vertex in the graph represents a person in online social network and edges between vertexes represent association or interaction between them. The problem of link prediction can be considered as a supervised learning model where we use the link information of training dataset to train the learning model. From the trained model, prediction of links can be made. Many researchers have used learning based algorithm to predict links in online social network. Hasan [9] identified set of features and shows effectiveness of features in link prediction method. They have compared different classes of supervised learning algorithm in terms of their prediction accuracy. In [10] authors have considered weighted network using supervised learning model. In their work, comparison of the link prediction model in co authorship network has been done which shows better results in supervised learning model comparatively to unsupervised model. Similarly in [11] authors have used classification model in the healthcare domain where model can predict future association among physician and shows good results in experiment. As common neighbour is one of simplest similarity measures between two nodes, the authors in [12] have used common neighbours between nodes and proposed, a probabilistic model using the naive bayes classifier where by identifying different roles of common neighbours and by giving different weights to them , the proposed model outperform in the experiment results. There exists no. Of classification model for supervised learning such as SVM, k nearest neighbour, decision tree. In [9] authors have used the co-authorship dataset and have used bagging and support vector machine for learning model. Also the regression model is used in [13] for the link prediction. For predicting links in the network, this problem also being solved using feature based link prediction model, where any popular supervised classification techniques can be employ [14] and the major challenge is to select the set of features.In this feature based learning model, topological features of the graph is considered mostly by the researchers [15]. There are many topology based features like node based, path based features which being considered for feature based learning based model. Though many work has included topological features, but to the best of our knowledge, lesser have given importance to centrality based features and have used PCA on these features. In our proposed methodology we have included centrality features in the feature vector and further

used dimension reduction technique, PCA (principal component analysis) and oversampling of dataset to improve the Link prediction performance.

3. Methodology

The proposed approach is using dimension reduction technique which reduce the high dimension features to low dimension features. The proposed methodology for feature based link prediction using PCA is represented in the Figure 2 and steps are explained as follows:

3.1. **Steps in proposed approach:**

Step 1: Extract features from the dataset: As our proposed approach is based on feature based learning model, the selection of features is an important task. In our proposed method, following are the components of features vector:

1) Common Neighbours(CN) :Common neighbours metric is based on the thought that if two nodes i and j have many common neighbour nodes, the probability to have link in the future is more.

2) Resource allocation index (RAI): It is motivated by a resource allocation process and measures how much resource is transmitted between a and b.. Therefore, the similarity of node a and node b can be defined as the sum of the inverse of the degree of each of the common neighbor between a and b.

3) Betweeness Centrality (BC): Betweeness centrality, determines of the degree to which a given node is in the shortest paths between the other nodes in the graph. A node (a) has high Betweeness if the shortest paths between many pairs of the other nodes in the graph pass through that node (a).

4) Closeness centrality (CLC): It identifies the most significant nodes in the network. For node a, CLC is defined as the ratio of the total number of nodes(N) in the graph minus one to the sum of the shortest distances of the node a to every other node in the graph.

5) Degree Centrality(DC) : Degree centrality is defined as the no. of connections of the node in a network graph G(N,E) .The Degree centrality is normalized by dividing is by the maximum degree in graph and defined as: DC(a)(Degree Centrality of node a)=K(a)/(N-1) K(a) is the degree of the node a and (N-1) maximum degree in a graph.

Step 2: Balance the Data using oversampling technique: By Using the extracted features of the network, the links between the nodes will be identifies a F link or NF link. Before performing classification we have checked for any missing values and balance the data using Oversampling of the data. For the balancing of the dataset oversampling or under sampling can be done. For our work we have employed oversampling of the data , by which the the resultant dataset have become balanced.

Step 3: Apply Dimension reduction technique, Principal Component Analysis (PCA) on the Resultant Feature vector which provides relevant and uncorrelated dataset. The PCA is a procedure that converts the correlated attributes into linearly uncorrelated attributes.

Step 4: Perform Classification using Principal components (PCs)

Step 5: Evaluate the model using Accuracy, recall, precision and F-measure.

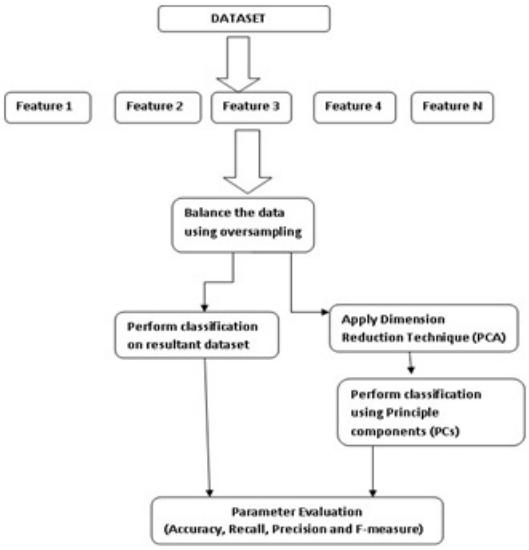

FIGURE 2. Proposed approach for feature based link prediction using PCA

3.2. **Evaluation metrics that we have used in our experiment:**

1. Accuracy: It is the performance metrics to measure the accuracy of the model. As defines in [16], it is the ratio of total no. of correct Predictions to the total no. of samples or predictions

Accuracy= Correct Prediction (True Positive+ True Negative)/Total No. of Samples

2. Precision: Precision talks about how precise and accurate model is out of total predictive positive.It is define as no. of true positive samples divided by total no. of predictive positive samples.

Precision: True positive samples/ Total Predictive Positive samples

3. Recall: Recall metrics is the ratio of True positive samples to the total actual positive samples.

Recall : True positive samples/ Total Actual Positive samples

4. F-measure: F -measure is the function of precision and recall and it is defined as:

F-measure = 2 * (Precision * recall) / (Precision + recall)

4. EXPERIMENT RESULTS AND DISCUSSION

We evaluate our model on karate club dataset; the network graph of the dataset is shown in the Figure 3.

In the Zachary karate club dataset, nodes are 34 and edges are 78. The node is represented as members of club and the relationship or link between the club members is represented by F link and those members who don't have any link is represented by NF link. In our proposed approach we have applied dimension reduction technique on feature vector which gives us principal components. Using these principle components we have

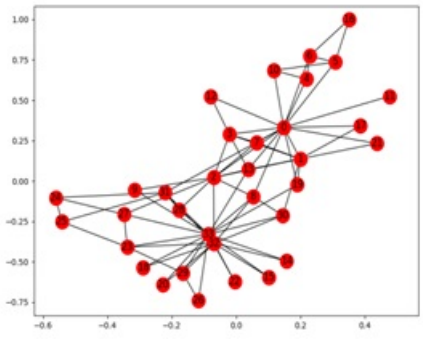

FIGURE 3. Network graph of karate club dataset, nodes=34 and F links=78

TABLE 1. ACCURACY USING PCA AND WITHOUT USING PCA

Classifier/Parameter for Evaluation	KNN	Decision Tree	Naive Bayes	SVM
Using PCA	0.9343	0.8216	0.7559	0.8779
Without using PCA	0.8967	0.7887	0.7512	0.8545

applied the classification using K nearest neighbour, Naive bayes (NB), SVM (support vector machine) and Decision Tree. In Table I and Figure 4, the accuracy of proposed approach using PCA has been compared with classification of links without using PCA. The results are better when we have applied the dimension reduction technique.

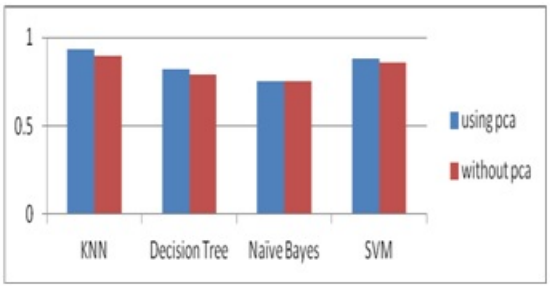

FIGURE 4. Accuracy using PCA and without using PCA

Other parameters that we have used for evaluation of our proposed approach is Precision, recall and F-measure. The results of classification of links with PCA and without using PCA are shown in Table II and Figure 5. F-measure which is balanced of recall and precision is better using PCA than without using PCA in all the cases.

TABLE 2. RECALL, PRECISION, F-MEASURE USING PCA AND WITHOUT USING PCA

Classifier/Parameter for Evaluation	KNN	Decision Tree	Naive Bayes	SVM
Recall using PCA	0.88	0.85	0.81	0.86
Recall without PCA	0.82	0.85	0.84	0.79
Precision using PCA	1	0.77	0.66	0.92
Precision without PCA	1	0.7	0.61	0.95
F- measure using PCA	0.93	0.80	0.72	0.88
F-measure without PCA	0.90	0.76	0.70	0.86

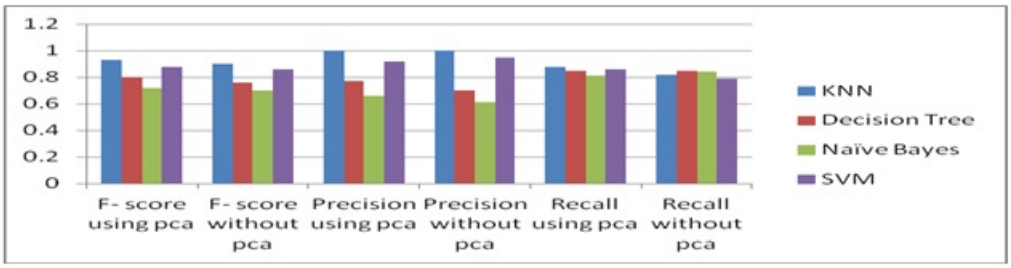

FIGURE 5. F-score, Recall and Precision

5. CONCLUSION

Dimension reduction technique allows to reduce the high dimension data to low dimension data and gives the uncorrelated dataset. Using one of the most popular techniques of dimension reduction i.e. PCA (Principle component analysis), we have proposed a feature based learning model. In our approach we have balance the data using oversampling and then applied the Principal component analysis on the extracted features which is included Common neighbour, Resource allocation index and centrality based features (Degree centrality, Closeness centrality and Betweeness centrality). The results depict the performance (Accuracy and F-measure) of classification based link prediction model is improved after applying PCA.

REFERENCES

[1] Huang, Z, Dajun Zeng, D. A link prediction approach to anomalous email detection. In: IEEE International Conference on Systems, Man and Cybernetics, San Diego, CA, 2006, 1131?1136.

[2] Folino, F. and Pizzuti, C., Link prediction approaches for disease networks. In International Conference on Information Technology in Bio-and Medical Informatics. Springer, Berlin, Heidelberg, 2012, 99-108.

[3] Esslimani, I., Brun, A. and Boyer, A., Densifying a behavioral recommender system by social networks link prediction methods. Social Network Anal. Mining, 1(3)(2011), 159-172.

[4] Chen, H., Li, X. and Huang, Z., Link prediction approach to collaborative filtering. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05), IEEE, 2005, 141-142. .

[5] Lü, L., Jin, C.H. and Zhou, T., Similarity index based on local paths for link prediction of complex networks. Phys. Rev. E, 80(4)(2009), 046122.

[6] Liu, W. and Lü, L., Link prediction based on local random walk. Europhys. Lett. 89(5)(2010), 58007.

[7] Benchettara, N., Kanawati, R. and Rouveirol, C., Supervised machine learning applied to link prediction in bipartite social networks. In 2010 International Conference on Advances in Social Networks Analysis and Mining. IEEE 2010, 326-330.

[8] Liben-Nowell, David, and Kleinberg, Jon. The Link Prediction Problem for Social Networks. J. Amer. Soc. Inf. Sci. Technol. 58(7)(2007), 1019-1031

[9] Al Hasan, M., Chaoji, V., Salem, S. and Zaki, M., April. Link prediction using supervised learning. In SDM06: workshop on link analysis, counter-terrorism and security, 2006.

[10] De Sa, H.R. and Prudencio, R.B., Supervised link prediction in weighted networks. In The 2011 international joint conference on neural networks, IEEE, 2011, 2281-2288.

[11] Almansoori, W., Gao, S., Jarada, T.N., Elsheikh, A.M., Murshed, A.N., Jida, J., Alhajj, R. and Rokne, J., Link prediction and classification in social networks and its application in healthcare and systems biology. Network Modeling Analysis in Health Informatics and Bioinformatics, 1(1-2)(2012), 27-36.

[12] Liu, Z., Zhang, Q.M., Lü, L. and Zhou, T., Link prediction in complex networks: A local nave Bayes model. Europhys. Lett. 96(4)(2011), 48007.

[13] O'Madadhain, J., Hutchins, J. and Smyth, P., Prediction and ranking algorithms for event-based network data. ACM SIGKDD explorations newsletter, 7(2)(2005), 23-30.

[14] Al Hasan, M. and Zaki, M.J., A survey of link prediction in social networks. In Social network data analytics. Springer, Boston, MA. 2011. 243-275.

[15] Kashima, H. and Abe, N., A parameterized probabilistic model of network evolution for supervised link prediction. In Sixth International Conference on Data Mining (ICDM'06). IEEE. 2006, 340-349.

[16] Fawcett, T., An introduction to ROC analysis. Pattern Recognition Letters, 27(8)(2006), 861-874.

[17] Liu, H., Hu, Z., Haddadi, H. and Tian, H., Hidden link prediction based on node centrality and weak ties. Europhys. Lett. 101(1)(2013), 18004.

[18] Freeman, L.C., Centrality in social networks conceptual clarification. Social networks, 1(3)(1978), 215-239.

[19] Sabidussi G., The centrality of a graph, Psychometrika 31(4)(1966), 581-603.

[20] Yao, L., Wang, L., Pan, L. and Yao, K., Link prediction based on common-neighbors for dynamic social network. Proc. Computer Sci. 83(2016), 82-89.