ISSN 2047-3338

Link Prediction in Social Networks Using Fuzzy-based SVM

Sara Esmaeillou[1], Ali Soleimani[2] and Ramin Karimi[3]
[1,2,3]Islamic Azad University, Malard, Tehran, Iran

***Abstract*— Link prediction can be used in many cases, for instance, detecting and identifying convicts and criminals, which requires high accuracy and fault cost in this context is very high Thus prediction might increase probability of detecting such groups. Researches in the context of link prediction in social networks have mostly focused on conventional methods which are based on indices. In this paper, link prediction using SVM and Fuzzy is proposed. In the proposed method, Fuzzy and support vector machine are used simultaneously. Main idea of the proposed method is to classify samples using SVM and fuzzy membership functions each sample which has the highest matching regarding membership functions of each class belongs to that class. Sample fracture curves are obtained from SVM algorithm and then this curve is used in membership functions of Fuzzy algorithm.**

***Index Terms*— Social Networks, Link Prediction and FSVM**

I. INTRODUCTION

SOCIAL network is a social structure which is comprised of nodes (which are usually individual or organizational) that are connected through one or more dependencies, For example: prices, inspirations, financial ideas and exchanges, friends, family, trade, web links, epidemiology or airlines. Resulting structures are very complicated. Social network analysis considers social relations with vertices and edges. Vertices are individuals inside the network and edges are relations among these individuals. Several types of edges can exist among vertices [11]. Results of different researches show that capacity of social networks can be used at many individual and social levels for identifying problems and solving them, establishing social relations, administrating associational affairs and guiding people towards reaching their goals [29].

For instance, results of studies in tourism policy shows that social networks are effective in attracting foreign tourists through affecting different behavioral variables and these networks can be used to establish trust and reduce risk of users' decision about a particular tourism destination. In this paper, fuzzy-based SVM is proposed for link prediction in social networks and the obtained results are compared with those obtained from SVM method. In section II, previous studies in this context are reviewed. Section III presents SVM. Fuzzy method is presented in section IV and the proposed method is described in section V. Evaluation measures in link prediction, employed datasets and indices used in link prediction are presented in sections VI, VII and VIII, respectively. Section IX describes link prediction using FSVM. In section X, results obtained from methods based on SVM and FSVM are presented and section XI concludes the paper.

II. LITERATURE REVIEW

Cyberspace provides the possibility of forming new communities for users. Since Tunis and his effort for defining two types of human rally, community against society, all researchers in the field of social and cultural sciences, have proposed being face to face <> limited number <and> emotional relationships and not intellectual relations, as basic features of a community [24]. Different link prediction methods are categorized into three classes [33], [1]:

First class are the approaches based on similarity measures. These approaches use structural network graph features for their prediction. For example, length of shortest path between two nodes or number of shared neighbors between two neighbors and similarity measures [7]. Jaccuard coefficient method is similar to number of shared neighbors method, with this difference that in this method, pair of nodes which have high number of shared neighbors and low number of unshared neighbors are more similar [8]. In another method, link prediction is performed based on nodes' degree [4]. In this method, it is more probable that two nodes with higher degrees may communicate with each other in future [4], [6].

Second class methods are based on highest probability. In these methods, not only structure of the network is investigated but also its rules and features are extracted. However, in this class, special features of the network are considered but these methods are time consuming and can be performed on networks with low number of nodes [5], [17].

Third class methods are based on statistics. In these methods, statistical models and associated distributions are used for link prediction [1].

III. SUPPORT VECTOR MACHINE (SVM)

SVM can be used for applications like separating and classifying data. This algorithm can divide input data into two categories after training step, thus it has unique simplicity compared to intelligent algorithms like neural networks. But since it cannot divide into more than two categories, it is less used in cases where data should be divided to more than two categories. If data in a given space cannot be divided with one line. Eq. 1 is the meta-page data classifier. Eq. 2, Eq. 3 are meta-parallel page formulas based on maximum margin condition. As shown in Fig. 2, distance between two margin pages is [15], [9]:

$$w' x - b=0 \qquad (1)$$

$$w' x - b=-1 \qquad (2)$$

$$w' x - b=1 \qquad (3)$$

b is a vector, normal to the separator line and Ws are input variables. In the above equations, x can be decomposed linearly and based on Eq. 4 and Eq. 5, it lies in either class 1 or 2 [22].

$$w' x - b \text{ £ } <=1 \qquad (4)$$

$$w' x - b \text{ £ } >=1 \qquad (5)$$

IV. FUZZY

Fuzzy concept is an approach which resolves such problems because fuzzy rules are developed for modelling uncertainty and obscurity. Using fuzzy concepts increases interpretability and decreases sensitivity of rules to data. When an expert sees fuzzy terms at the output of the algorithm, can better understand the rule, because meaning of each fuzzy terms has originated from ambiguity of conversation which human can easily understand [2]. Fuzzy logic is based on fuzzy sets theory. This theory is a generalization of classic sets theory in mathematics. In classic sets theory, an element is either a member of the set or it is not. In fact, membership of elements follows a binary pattern; but fuzzy sets theory expands this concept and suggests graded membership. This way, an element can be a member of a set to some degrees and not completely [25]. For instance, “Mr. A is 70% member of adults’ community” is true in fuzzy sets theory. In this theory, membership of set’s members are determined using function u(x), Where x is a specific member and u is a fuzzy function which determines membership degree of x in the associated set and its value is between 0 and 1 [9].

In other words, (u)x is a mapping from x to possible values between 0 and 1. (u)x might be a set of discrete or continuous values[16]. When u comprises only a number of discrete values between 0 and 1, it might include 0.3, 0.5, 0.7 and 1; but when values of u are continuous, a curve of decimal numbers between 0 and 1 is formed. Fuzzy logic can be employed through rules which are called fuzzy operators. These rules are defined based on the following model:

IF variable IS set THEN action

Fuzzy logic is used along with other algorithms for optimizing algorithm process due to its high accuracy and in many articles it has been integrated with other methods as a basis for classification.

V. THE PROPOSED METHOD

In this article, flickr social network data which has 80513 nodes are investigated. In this method, first, available data are pre-processes, and nodes with degrees 0 and 1 are eliminated. After preprocessing, number of nodes and edges are reduced, then neighborhood set for each user is specified. And finally a matrix comprising neighbors is obtained for users. In the next step, appropriate features are obtained to use in next steps of the social network also train and test data are separated and each group is classified by the proposed algorithm.

VI. EVALUATION INDICES IN LINK PREDICTION

Wherever graph G(V, E) is considered where v is a set of nodes and E is a set of edges. Usually it is not important to know if the links are lost or what would be the links in future, because in practice this is not feasible and one can only study validity of results; therefore in order to test validity, observed links, E are divided into two groups randomly: train and test sets [13], [24]. Training set is used for testing and none of its data can be used for prediction, however information of training data are specified [17], [18].

Total data is the union of test and train data. Advantage of this random validation method is that number of iterations does not affect training data but in this method some links may not appear and some other links may appear at some other times and this might cause some statistical problems which can be resolved through employing k-fold cross-validation in which observed links are categorized into k random sub-sets and it is repeated k times. Using this method, all links are used for validation and each link is used for real prediction [20], [21]. In this case, the higher is K, more links are studied and computations are increased. The most appropriate case is 10-fold cross validation which results in good time and efficiency. Confusion matrix which presented in Table I, shows true and false predictions [19].

TABLE I: CONFUSION MATRIX: INCLUDING INDICES FOR CALCULATING PREDICTION RESULTS

Symbol	Prediction out come/ True State	
Φ	True Positives (TP)	hits
B	False Positives (FP)	false alarms
H	False Negatives (FN)	misses
m	True Negatives (TN)	correctly rejected

TABLE II:
EVALUATION MEASURES

Symbol		
Φ	Sensitivity (true positive rate, recall)	TPR=TP/(TN+TP)
B	Accuracy	TNR=(TN+TP)/(TN+FP+FN+TP)
H	Specificity (true negative rate)	TNR=TN/(TN+FP)
m	Positive predictive value (precision)	PPV=TP/(TP+FP)
σ	FMeasure	F=2. (precision .recall)/(precision+recall)
j	AUC	AUC=(0.5n"+n')/n

TABLE III.
PROPERTIES OF DATA GATHERED ON FLICKR

Symbol		Number
Φ	Categories (Number of groups)	195
B	Nodes	80،513
H	Links	5،899،882
m	Network density	1.8 ×10−3
σ	Maximum degree	5،706
j	Average degree	146
J	Clustering coefficient	0.61

According to Table I, true positive rate=Sensitivity=Recall is as follows:

$$TPR = \frac{\mathrm{TP}}{TN+TP} \qquad (6)$$

Accuracy is as follows (7)

$$ACC = \frac{\mathrm{TP+TN}}{TN+Fp+FN+TP} \qquad (8)$$

Positive predictive value or precision is also obtained as below:

$$PPV = \frac{\mathrm{TP}}{TP+FP} \qquad (9)$$

There are two other standard indices for validating link prediction algorithms including:
AUC is the area under ROC curve and it is calculated through independent comparisons [28], [32], [35], if n' times false link have higher score and n'' times false links and links which do not exist have the same score, AUC is obtained as below:

$$\mathrm{AUC} = \frac{0.5\mathrm{n}'' + \mathrm{n}'}{\mathrm{n}} \quad [34]$$

AUC gives score of all unobserved links. AUC value is interpreted as probability of selecting lost links randomly or train data links which have higher score compared to random selection of links which do not exist.
Precision which gives score of unobserved links.
F-Measure which considers both validity and recall for calculating precision and can be interpreted as a weighted measure of validity and recall. This weight is shown with α and it is usually considered 1. Value of F-measure is between 0 and 1 and the closer it is to 1, it shows better efficiency in classification [27].

VII. EMPLOYED DATA SETS

In this paper, data available in Flickr social network is used. These test data include images shared in 30/07/2010. These data include friendships and comments (who has commented for which image) among a set of users. Connected groups can be considered as tag for identifying society. Properties of these data are given in Table III.

VIII. INDICES USED IN LINK PREDICTION

Indices used in the proposed method are given in the following Table IV.

TABLE IV.
INDICES USED IN LINK PREDICTION

Symbol	Quantity	Conversion from Gaussian and CGS EMU to SI [a]
Φ	Common Neighbors [26]	$s_{xy} = \lvert N_{(x)} \cap N_{(y)} \rvert$
B	Jaccard Coefficients [23],	$s_{xy} = \left\lvert \frac{N_{(x)} \cap N_{(y)}}{N_{(x)} \cup N_{(y)}} \right\rvert$
H	Preferential Attachment [3]	$s_{xy} = \lvert K_x \times K_y \rvert$

m	Adamic Adar Coefficient [16]	$s_{xy} = \left\lvert \sum_{z\in N_{(x)}\cap N_{(y)}} \frac{1}{\log k_z} \right\rvert$
M	Resource Allocation Index [25]	$s_{xy} = \left\lvert \sum_{z\in N_{(x)}\cap N_{(y)}} \frac{1}{k_z} \right\rvert$
4πM	Salton Index [10]	$s_{xy} = \frac{\lvert N_{(x)} \cap N_{(y)} \rvert}{\sqrt{K_x \times K_y}}$
σ	Sorensen Index [30]	$s_{xy} = \frac{2\lvert N_{(x)} \cap N_{(y)} \rvert}{K_x + K_y}$
j	Hub Promoted Index [33]	$s_{xy} = \frac{2\lvert N_{(x)} \cap N_{(y)} \rvert}{\min\{K_x + K_y\}}$
J	Hub Depressed Index [2]	$s_{xy} = \frac{2\lvert N_{(x)} \cap N_{(y)} \rvert}{\max\{K_x + K_y\}}$
χ, κ	Leicht Hol me New man Index [7]	$s_{xy} = \frac{N_{(x)} \cap N_{(y)}}{K_x \times K_y}$

TABLE V
LINK PREDICTION RESULTS USING SVM AND FSVM

Symbol		SVM (C=0.1)	Proposed Method
Φ	Accuracy	97	97.88
B	F-measure	0.96	0.97
H	Precision	0.96	0.98
m	Recall	0.97	0.97
M	Auc	0.59	0.89

IX. LINK PREDICTION USING FSVM

In this method fuzzy and SVM approaches are employed simultaneously and link prediction problem is proposed as a two-class problem such that available edges are tagged as 1 and edges which do not exist are tagged as 0, and these two classes are known as positive and negative classes [9], [15]. Our purpose is to predict edges which will turn to 1 among those edges which are tagged 0 and vice versa.

In the proposed method, instead of using a SVM algorithm, a hybrid algorithm as fuzzy support vector machine is used. Main idea of the proposed method is the way samples are classified using SVM and membership functions of fuzzy algorithm, in other words this method classifies data which are beyond classification scope based on membership functions such that distribution error of each class is decreased. Therefore, each sample with highest matching with membership functions of each class belongs to that class. In this method, bilateral weighted error is used to map samples of each class to membership functions, such that weighted error function is calculated for each positive and negative class and membership functions are calculated for all samples and each sample whose membership function has a smaller error belongs to that class.

In this method, samples are considered as set of points like SVM method. In order to reduce error and use fuzzy concept in classifying samples, two points are considered, i.e. each positive sample is considered as good and bad, thus number of points is considered N2 instead of N. for instance, each point in training set is considered as N1k={xk,yk}, but in the proposed method each point is considered as N1K={-m_K1,1,m_K} , {x_k,-1,x_k} where x_k represents kth input vector and y_k represents kth output vector and m_k represents kth member of y_k class. Thus classification problem is reviewed as below:

$$min_{w,b.ŋ,\varepsilon_k}£(w,b,n_k,\varepsilon_k)$$
$$= \frac{1}{2}ww^T + c\sum_{k=1}^{n}(m_k\varepsilon_{k+}\ (1-m_k)n_k)$$

Subject to :

$$w^T\phi(x_k) + b \geq 1 - \varepsilon_k \qquad \text{k=1,2,…,N}$$
$$w^T\phi(x_k) + b \geq 1 + n_k, \qquad \text{k=1,2,…,N}$$
$$\varepsilon_k \geq 0, \qquad \text{k=1,2,…,N}$$
$$n_k \geq 0, \qquad \text{k=1,2,…,N}$$

In addition, membership functions or positive class errors are calculated using the following equations:

$$min_{w,b.ŋ,\varepsilon_k}£(w,b,n_k,\varepsilon_k) = \tfrac{1}{2}ww^T + c\textstyle\sum_{k=1}^{n}(m_k\varepsilon_{k+}\ (1 - mk)nk)$$

Subject to :

$$w^T\phi(x_k) + b \geq 1 - \varepsilon_k \qquad \text{k=1,2,…,N}$$

$$w^T\phi(x_k) + b \geq 1 + n_k \qquad \text{k=1,2,…,N}$$

Error Term =

$$\begin{cases} (1-m_k)(w^T\phi(x_k) + b + 1) & ,w^T\phi(x_k) + b \geq 1\ ) \quad (10) \\ (1-m_k)(w^T\phi(x_k) + b + 1) + m_k(1 - w^T\phi(x_k) - b) & \\ \qquad ,-1 \leq w^T\phi(x_k) + b \leq 1 & , \\ m_k(1 - w^T\phi(x_k) - b), & w^T\phi(x_k) + b \leq 1 \end{cases}$$

Error Term=

$$\begin{cases} 0, & w^T\phi(x_k) + b \geq 1 (11) \\ (1 - w^T\phi(x_k) - b), & -1 \leq w^T\phi(x_k) + b \leq 1 \\ (1 - w^T\phi(x_k) - b), & w^T\phi(x_k) + b \leq 1 \end{cases}$$

Error Term=

$$\begin{cases} 0, & w^T\phi(x_k) + b \geq 1\ (12) \\ m_k(1 - w^T\phi(x_k) - b), & -1 \leq w^T\phi(x_k) + b \leq 1 \\ m_k(1 - w^T\phi(x_k) - b), & w^T\phi(x_k) + b \leq 1 \end{cases}$$

Where the difference between SVM algorithm and the proposed algorithm is the symmetric or bilateral weighted error function or membership function. Since in this research, number of samples are doubled and for each sample, its membership is calculated with a function for that class. For this reason, in order to calculate algorithm's error, membership of that sample is also involved. Accordingly, instead of calculating error, weighted error is used [19,31]. According to{x_k,1, m_k} ,{ x_k,-1, 1-m_k, xk }belongs to class 1 with value mk and xk belongs to the opposite class with value mk-1, otherwise sample is negative and it is labeled as negative samples.

X. RESULTS OBTAINED USING METHODS BASED ON SVM AND FSVM

Although Fuzzy classification is one of the high power of categories, but is usually less than the others. fsvm is a classifier that is used in some cases that the test samples have different degree of importance , such as the social network , fsvm is svm very developed, The proposed method is almost an optimal way for link prediction in social networks that lead to higher standards of accuracy, Recall, F-measure, Precision and specialy Auc. Results of link prediction for social networks using svm and fsvm are represented in Table V.

XI. CONCLUSION

One of the link prediction approaches in social networks is based on machine learning which is performed by feature extraction and sample learning. In this paper, this approach is employed by combining fuzzy algorithm and SVM to propose an efficient algorithm for link prediction. And the results show that the proposed method has improved significantly due to efficient separation of data and using strong learner.

REFERENCES

[1] A. Clauset, C. Moore, M. E. Newman, "Hierarchical structure and the prediction of missing links in networks", 2008.

[2] Ahmad KhaliliJafarabad, "Fuzzy models for link prediction in social network", 2013.

[3] A.-L. Barabلsi, R. Albert, "Emergence of scaling in random networks, Science", 1999.

[4] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis, "Manolopoulos, Friendlink: Link Prediction in Social Networks via Bounded Local Path Traversal", 2011.

[5] Catherine A. Bliss, Morgan R. Frank, Christopher M. Danforth, Peter Sheridan Dodds, "An Evolutionary Algorithm Approach to Link Prediction in Dynamic Social Networks", 2014.

[6] Charu C. Aggarwal, Social Network Data Analytics, Springer, New York,

[7] E.A. Leicht, P. Holme, M.E.J. Newman, Vertex, "similarity in networks", 2006.

[8] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, A.-L. Barbasi, "Community structure in social and biological networks", 2002.

[9] Evaggelos Spyrou, Giorgos Stamou, Yannis Avrithis and Stefanos Kollias, "Fuzzy Support Vector Machines For Image Classification Fusing MPEG-7 Visual Descriptors", 2011.

[10] G. Salton, M.J. McGill," Introduction to Modern Information Retrieval", 1983.

[11] Hasan, Mohammad A., and Chaoji, Vineet, and Salem, Saeed and Zaki, Mohammed. Link Prediction using Supervised Learning. Proceedings of SDM Workshop of Link Analysis, Counterterrorism and Security. 2006.

[12] H. Kashima and N. Abe, "A parameterized probabilistic model of network evolution for supervised link prediction", 2006.

[13] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: recommending people on social networking sites", pp. 201-210, 2009.

[14] Kleinberg, Jon M, "Navigation in a small world", 2000.

[15] Kobe University, Rokkodai, Nada, Kobe, Japan" Fuzzy support vector machines for multilabel classification", pp. 2110–2117, 2005.

[16] L. Adamic and E. Adar, "How to search a social network. Social Networks", pp. 187–203, 2005.

[17] Lars Backstrom, and Jure Leskovec, "Supervised Random Walks: Predicting and Recommending Links in Social Networks", 2011.

[18] Liben-Nowell, David, and Kleinberg, "The Link Prediction Problem for Social Networks", Journal of the American Society for Information Science and Technology, pp. 1019-1031, 2007.

[19] Linyuan Lü, Tao Zhou, "Link prediction in complex networks: A survey", Elsevier, pp. 115–117, 2011.

[20] Michael Fire, Lena Tenenboim, Ofrit Lesser, Rami Puzis, Lior Rokach and Yuval Elovici, "Link Prediction in Social Networks using Computationally Efficient Topological Features", IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, 2011.

[21] Mohammad Al Hasan, and Mohammed J. Zaki, A Survey Of Link Prediction In Social Networks, Springer Science, Business Media, LLC, 2011.

[22] Omid Naghash Almasi1, Hamed Sadeghi Gooqeri, Behnam Soleimanian A. and Wan Mei Tang, "A New Fuzzy Membership Assignment Approach for Fuzzy SVM Based on Adaptive PSO in Classification Problems", Journal of mathematics and computer science, Vol. 14, pp. 171-182, 2015.

[23] P. Jaccard, Stude comparative de la distribution florale dans une portion des alpes et des jura, Bull. Soc. Vaud. Sci. Nat. 37 (1901) 547.

[24] Purnamrita Sarkar and AndrewW. Moore, "Random Walks In Social Networks And Their Applications, 2011.

[25] Q. Ou, Y.-D. Jin, T. Zhou, B.-H. Wang, B.-Q. Yin," Power-law strength-degree correlation from resource-allocation dynamics on weighted networks", 2007.

[26] Taoufik Guernine and Kacem Zeroual" New fuzzy multi-class method to train SVM classifier,2011

[27] R. K. Sivagaminathan and S.Ramakrishnan, "Ahybrid approach for feature subset selection using neural networks and ant colony optimization, 2007.

[28] Teshome Feyessa, Marwan Bikdash and Gary Lebby, "Node-pair Feature Extraction for Link Prediction", IEEE International Conference on Privacy, Security, Risk, 2011.

[29] Thelwall M. Social network sites: users and uses, Advances in computers, Vol. 76, pp. 23-26, 2009.

[30] T. Sorensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons", 1948.

[31] T. Zhou, L. Lü, Y.-C. Zhang, "Predicting missing links via local information", 2009.

[32] Wang, Chao, and Satuluri, Venu, and Parthasarathy, Srinivasan, “Local Probabilistic Models for Link Prediction”, 2007.

[33] WANG Peng, XU BaoWen, WU YuRong and ZHOU XiaoYu, “Link Prediction in Social Networks:the State-of-the-Art”, 2015.

[34] William Cukierski, Benjamin Hamner, Bo Yang” Graph-based Features for Supervised Link Prediction”, 2011.

[35] Xiaoqing Gu, Tongguang Ni, and Hongyuan Wang “New Fuzzy Support Vector Machine for the Class Imbalance Problem in Medical Datasets Classification", 2014.

[36] XufeiWang, Lei Tang, Huan Liu, LeiWang, “Learning with multi-resolution overlapping communities”, 2012.

Sara Esmaeillou is M.Sc graduated in the Department of Computer Engineering in Malard Branch, Islamic Azad University, Tehran, Iran. She received her B.S degree in Computer Engineering from Azad University and her research interests include Ecommerce, Social Network and Ad Hoc Networks.
Email: sara.esmaeillou@gmail.com, Phone Number: (+98-912) 2471778

Dr. Ali Soleimani, researcher and professor of university, holds a doctorate degree in computer science in Emerging Media from Colorado Technical University (Colorado, United State of America). He has over 20 years of experience in computer science. Dr. Soleimani’s researches blend education, computer science, and virtual worlds.
Email: a.soleimani.uni@aol.com, Phone Number: (+98)912-840-2831

Ramin Karimi is Assistant Professor in the Department of Computer Engineering in Malard Branch, Islamic Azad University, Tehran, Iran. He received Ph.D degree in Information Science, at Universiti Teknologi Malaysia, Johor, Malaysia in 2013. He received M.Sc degree in Computer Engineering from Iran University of Science and Technology in 2006 and his research interests include Vehicular Ad Hoc Networks, Mobile ad-hoc networks, Mobile Robots, Mobile Ecommerce, security and communication Networks.
Email: rakarimi1@gmail.com, Phone Number: (+98-912) 2382300