ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201812.00115

Edge Sign Prediction in Social Networks Based on Node Status and Similarity (Postprint)

Authors: Lu Zhigang, Ye Meili

Date: 2018-12-13T00:00:00+00:00

Abstract

Edge sign prediction mines sign-related implicit information based on network topology, aiming to reveal potential relationships between users. Node status and similarity can effectively reflect edge sign properties, providing a theoretical basis for improving prediction effectiveness. By investigating the strong correlation between these two factors and edge sign properties, a sign prediction model is constructed. First, the Prestige ranking algorithm is utilized to assess the social status of user nodes, while cosine similarity is concurrently employed to represent users' social preferences; then, the two factors are integrated on the basis of the logistic regression learning model to establish the edge sign prediction model LR-SN; finally, the stochastic gradient ascent algorithm is adopted for optimization during model training. Experimental results on three real-world network datasets demonstrate that, compared with existing benchmark methods, the LR-SN model achieves significantly improved sign prediction accuracy and exhibits certain generalizability, indicating that fusing local and global information can further enhance prediction performance.

Full Text

Preamble

Social Network Edge Sign Prediction Based on Node Status and Similarity

Lu Zhigang, Ye Meili
(College of Economics & Management, Shanghai Maritime University, Shanghai 201306, China)

Abstract: Edge sign prediction aims to mine sign-related implicit information from network topology to reveal potential relationships between users. Node status and similarity can effectively represent edge sign attributes, providing a theoretical basis for improving prediction performance. By investigating the

strong correlation between these two factors and edge sign attributes, we establish a sign prediction model. First, we use the Prestige ranking algorithm to evaluate the social status of user nodes, while employing cosine similarity to represent users' social preferences. Then, we fuse these two components based on a logistic regression learning model to establish the edge sign prediction model LR-SN. Finally, we adopt a stochastic gradient ascent algorithm for optimization during training. Experimental results on three real-world network datasets demonstrate that compared with existing baseline methods, the LR-SN model significantly improves sign prediction accuracy and exhibits certain generalizability, indicating that fusing local and global information can further enhance prediction effectiveness.

Keywords: edge sign prediction; node status; node similarity; logistic regression; stochastic gradient ascent algorithm

0 Introduction

Social networks are virtual spaces where people exchange opinions and share information, allowing users to label connected individuals as friends or enemies and express agreement or disagreement with others' statements and viewpoints. Therefore, social networks can be described as directed networks with positive or negative edge signs, where positive edges represent constructive relationships such as friendship, trust, or liking between two users, while negative edges represent destructive relationships such as hostility, suspicion, or dislike. Edge sign prediction in social networks involves extracting network structural information and user relationship data to predict unknown edge signs, revealing potential relationships like friends, strangers, or enemies.

Edge sign prediction holds significant research importance in machine learning, big data analytics, and decision-making. Investigating edge sign attributes helps understand fundamental network structural characteristics [1] and addresses problems such as personalized recommendation [2], public opinion analysis [3], and anomalous user detection [4]. This paper delves into edge sign attributes and proposes an efficient edge sign prediction model. We conduct extensive experiments on Epinion, Slashdot, and Wikipedia datasets, demonstrating the effectiveness of our model in sign prediction.

The main contributions are as follows:
a) We propose two quantification strategies for sign attributes, quantifying node status and similarity respectively.
b) Based on the logistic regression learning model, we fuse node status and similarity to establish the edge sign prediction model LR-SN, where node status quantifies sign attribute-related features from a global perspective, and node similarity embodies sign attributes from a local perspective.
c) To validate the effectiveness of the LR-SN model, we conduct multiple experiments on Epinion, Slashdot, and Wikipedia datasets, and elaborate in detail

on how different quantification strategies affect sign prediction accuracy.

1 Related Research

Research on social network edge signs originated from social psychology, initially explored by Heider et al. [5] from a psychological perspective on interaction patterns between positive and negative relationships in interpersonal communication. Subsequently, Cartwright and Harary [6] described social networks using graph theory as directed networks with positive and negative edge signs. With the rise of complex networks, edge sign prediction in social networks has gradually become a research hotspot.

Current methods for edge sign prediction can be broadly divided into two categories: local feature-based methods and global feature-based methods. Local feature-based methods utilize only neighborhood characteristics of nodes, such as node in-degree/out-degree [7], number of common neighbors [7], and node similarity [8-9], for edge sign prediction. Global feature-based methods expand the scope of feature extraction, quantifying network imbalance from a global perspective, typically employing extended structural balance theory [10], contextual information [11,12], node ranking [13], and other measures for sign prediction. Leskovec et al. [7] formalized the sign prediction problem by extracting two types of network structural information: node neighborhood features and 16 types of triadic relationship patterns based on sociological theories, then trained features using logistic regression to achieve edge sign prediction. Chiang et al. [10] proposed using extended structural balance theory with ordered long cycles for edge sign prediction, demonstrating that prediction accuracy improves effectively as cycle length increases from 3 to 5. This method extends the local metric approach of Leskovec et al. Symeonidis et al. [9] defined intra-cluster and inter-cluster similarity, then utilized recommendation algorithms for sign prediction. Shahriari and Jalili [13] introduced ranking algorithms into feature value computation, first ranking nodes in the network using various sorting algorithms, then calculating features based on these ranking values, thereby embodying edge sign attributes from a global perspective.

Local and global information in networks are closely related, but using only one of them for edge sign prediction is insufficient. To address this issue, this paper fuses local and global information by introducing node status and similarity quantification strategies based on the logistic regression learning model, thereby solving problems such as low prediction accuracy caused by network sparsity and insufficient utilization of local features.

2 Problem Formalization

We represent a social network as a directed graph denoted as $G(V, E, S)$, where $V = \{1, 2, ..., n\}$ represents the set of user nodes in the social network, $E = \{1, 2, ..., m\}$ represents the set of edges depicting relationships between user nodes, and $S = \{1, -1, 0\}$ represents edge signs. For nodes $i, j \in V$, if $(i, j) \in E$, then $s_{ij} \in S$. When $s_{ij} = 1$, it represents a positive relationship indicating trust, cooperation, friendliness, or support between nodes i and j; when $s_{ij} = -1$, it represents a negative relationship indicating distrust, hostility, dislike, or opposition between two nodes; and $s_{ij} = 0$ indicates that the interaction relationship between nodes i and j is unknown.

Using the above notation and definitions, we define the edge sign prediction problem in social networks as follows: Design a sign prediction framework $f(G, V, E, S)$ that predicts edge signs in the network by extracting network structural information combined with known user relationship data. That is, given a social network $G(V, E, S)$, the framework f predicts unknown positive or negative relationships: $f(G, V, E, S) \rightarrow s_{ij}$.

---

3 Node Status and Similarity Quantification

Empirical studies have found that edge sign attributes are closely related to node status and similarity [7-9], both of which can reflect users' popularity and social preferences in the network to some extent. Therefore, we design a sign prediction framework f using node status and similarity, and propose two quantification strategies for sign attributes.

3.1 Node Status

Differences in node status within a network can be used to label edge signs [14]. Specifically, a positive edge from i to j indicates that i' s status is higher than j' s, while a negative edge from i to j indicates that i' s status is lower than j' s. This hierarchical relationship is transitive [15]. Network topology can be used to evaluate the social status of user nodes. In social networks, positive edge in-degree helps elevate a node' s social status, while negative edge in-degree reduces it. Based on this, we propose a node status quantification strategy—Prestige—to evaluate users' social status.

Prestige [12] considers only the positive and negative edge in-degree of nodes in the network structure. If a node receives many positive edges from other nodes, it indicates high status and prestige in the network. Conversely, if a node receives many negative edges, its status and reputation are low. The Prestige value of node i (Pr_i) is calculated as follows:

$$Pr_i = \frac{IN_i^+ - IN_i^-}{IN_i^+ + IN_i^-}$$

where IN_i^+ and IN_i^- represent the positive and negative edge in-degree of node i, respectively. A higher Prestige value indicates greater prestige and status, making the node more easily trusted in the network. Conversely, a lower Prestige value suggests the node is less likely to be trusted by other nodes.

3.2 Node Similarity

Users with similar preferences tend to assign similar signs to edges in social networks [8]. Node similarity can roughly reflect users' social preferences in the network. For nodes related to the edge sign to be predicted, we can infer the likelihood of the source node assigning a positive or negative edge to the target node by calculating the average similarity between the source node' s neighbors and the target node' s neighbors. Intuitively, if user i has high similarity with users who have given positive edges to user j, then user i is likely to give a positive edge to user j. Conversely, if user i has high similarity with users who have given negative edges to user j, then user i is likely to give a negative edge to user j.

Let s_{ij} denote the edge sign from node i to node j, and N_i and N_j denote the neighbor nodes of i and j, respectively. We use cosine similarity to calculate similarity between user nodes, defined as follows:

$$Sim(i,k) = \frac{\sum_{p \in I_{ik}} r_{ip} \cdot r_{kp}}{\sqrt{\sum_{p \in I_{ik}} r_{ip}^2} \sqrt{\sum_{p \in I_{ik}} r_{kp}^2}}$$

Based on known social network topology information, we quantify the likelihood of user i giving a positive edge to user j through node similarity:

$$S_{ij}^+ = \frac{\sum_{k \in W^+} Sim(i,k)}{|W^+|}$$

where $W^+ = \{k | s_{kj} = 1\}$ represents the set of nodes that have positive edges with node j, and $|W^+|$ denotes the set size.

Conversely, the likelihood of user i giving a negative edge to user j is:

$$S_{ij}^- = \frac{\sum_{k \in W^-} Sim(i,k)}{|W^-|}$$

where $W^- = \{k | s_{kj} = -1\}$ represents the set of nodes that have negative edges with node j.

4 Edge Sign Prediction Models

Using node status and similarity as the foundation for model construction, we establish edge sign prediction models based on logistic regression, with corresponding quantification strategies.

4.1 Logistic Regression-Based Edge Sign Prediction (LR)

Logistic regression is a supervised statistical learning method that treats positive/negative relationship prediction in social networks as a binary classification problem. To apply this algorithm for sign prediction, we first construct a feature set related to edge sign attributes, then use this feature set as input to train a classifier for positive/negative relationship prediction. The specific form is as follows:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

where $x = (x_0, x_1, x_2, ...)$ represents the feature vector extracted from the social network, which should reflect edge sign attributes to a certain extent. $\theta = (\theta_0, \theta_1, \theta_2, ...)$ represents the weight vector assigned to each feature, estimated through maximum likelihood methods.

The probability that edge y to be predicted is positive is:

$$P(y = 1|x; \theta) = h_\theta(x)$$

The probability that edge y is negative is:

$$P(y = 0|x; \theta) = 1 - h_\theta(x)$$

Edge y has two possible values: 0 or 1. When $y = 1$, the predicted edge is positive; when $y = 0$, the predicted edge is negative. Generally, when $P(y = 1|x; \theta) > 0.5$, the predicted edge y is positive (value 1), and when $P(y = 1|x; \theta) < 0.5$, the predicted edge y is negative (value 0).

4.2 Node Status-Based Edge Sign Prediction Model (LR-S)

To simulate node status for edge sign prediction, we establish a node status-based edge sign prediction model LR-S by leveraging the strong correlation between node status and edge sign attributes. To emphasize the importance of node status in the prediction process, we propose four node status quantification features as the baseline. These features quantify each user node' s social status and optimism degree from a global perspective, largely reflecting edge sign attributes.

The node status quantification features include Rep_i, Opt_i, Rep_j, and Opt_j, representing the reputation value and optimism value of node i, and the reputation value and optimism value of node j, respectively. In the feature calculation process, we first use the node status quantification strategy—Prestige—to evaluate each user node' s social status, then calculate the reputation and optimism values of nodes related to edge sign prediction based on the obtained Prestige values.

The reputation value and optimism value of node i are defined as:

$$Rep_i = \frac{\sum_{k \in IN_i^+} Pr_k - \sum_{k \in IN_i^-} Pr_k}{|IN_i^+| + |IN_i^-|}$$

$$Opt_i = \frac{\sum_{k \in OUT_i^+} Pr_k - \sum_{k \in OUT_i^-} Pr_k}{|OUT_i^+| + |OUT_i^-|}$$

Similarly, the reputation value and optimism value of node j are defined as:

$$Rep_j = \frac{\sum_{k \in IN_j^+} Pr_k - \sum_{k \in IN_j^-} Pr_k}{|IN_j^+| + |IN_j^-|}$$

$$Opt_j = \frac{\sum_{k \in OUT_j^+} Pr_k - \sum_{k \in OUT_j^-} Pr_k}{|OUT_j^+| + |OUT_j^-|}$$

where Pr_k is the ranking score of node k, calculated using Equation (2). IN_i^+ and IN_i^- represent the sets of nodes that give positive and negative edges to node i, respectively. OUT_i^+ and OUT_i^- represent the sets of nodes that receive positive and negative edges from node i, respectively. Similar definitions apply to node j.

A node' s reputation value reflects its popularity in the network, measuring its acceptance within its social circle and its influence in society. Nodes with higher reputation values have greater status and influence, making other nodes more likely to give them positive edges. Conversely, the optimism value reflects a node' s tendency to give positive, friendly interactive relationships to other nodes in the network, indicating the node' s personality— "optimistic and friendly" to a certain extent. The higher the optimism value, the more likely the node is to give positive edges to other nodes.

The logical relationship between reputation value and optimism value is shown in Figure 1 [Figure 1: see original paper]. The reputation value only considers the set of nodes related to node i' s incoming edges and calculates the Prestige values of all nodes in this set. Conversely, the optimism value only considers the set of nodes related to node i' s outgoing edges and the Prestige values of all nodes in this set.

Based on the logistic regression sign prediction method, we achieve model extension by introducing four node status quantification features as the feature set input. Thus, the LR-S model can be defined as:

$$P(y=1|x;\theta)=\frac{1}{1+e^{-(\theta_0+\theta_1 Rep_i+\theta_2 Opt_i+\theta_3 Rep_j+\theta_4 Opt_j)}}$$

4.3 Node Status and Similarity-Based Edge Sign Prediction Model (LR-SN)

Although node status can quantify sign attribute-related features from a global perspective, it still has limitations. First, when the social status difference between two users is small, prediction accuracy will be affected. Second, node status focuses on global information in the network while neglecting local information. Empirical studies have found that when network clustering coefficient is high, local information can achieve higher prediction accuracy than global information. Therefore, when establishing a node status-based edge sign prediction model, we must consider not only the social status difference between two users but also incorporate node similarity that reflects local information to overcome its limitations in sign prediction. Thus, we propose the node status and similarity-based edge sign prediction model LR-SN.

Compared with LR-S, this model innovatively proposes four additional node similarity quantification features. These features quantify each user' s attributes and preferences from a local perspective, further embodying edge sign attributes.

The node similarity quantification features include source node positive similarity, source node negative similarity, destination node positive similarity, and destination node negative similarity, defined as follows:

a) **Source Node Positive Similarity** $S_{ij}^{+,out}$: The average similarity between node i and nodes that give positive edges to j. The higher the $S_{ij}^{+,out}$ value, the greater the likelihood that the edge from node i to node j is positive. $S_{ij}^{+,out}$ is calculated as:

$$S_{ij}^{+,out}=\frac{\sum_{k\in W_{out}^{+}} Sim_{out}(i,k)}{|W_{out}^{+}|}$$

where $W_{out}^{+}=\{k|s_{kj}=1\}$ is the set of nodes that give positive edges to node j, and $Sim_{out}(i,k)$ is the similarity between node i and node k that gives positive edges to j, calculated as:

$$Sim_{out}(i,k)=\frac{\sum_{p\in I_{ik}^{out}} r_{ip}\cdot r_{kp}}{\sqrt{\sum_{p\in I_{ik}^{out}} r_{ip}^2}\sqrt{\sum_{p\in I_{ik}^{out}} r_{kp}^2}}$$

where r_{ip} and r_{kp} are the edge signs from nodes i and k to node p, respectively, and I_{ik}^{out} is the set of nodes that both i and k point to.

b) **Source Node Negative Similarity** $S_{ij}^{-,out}$: The average similarity between node i and nodes that give negative edges to j. The higher the $S_{ij}^{-,out}$ value, the greater the probability that the edge from node i to node j is negative. $S_{ij}^{-,out}$ is defined as:

$$S_{ij}^{-,out} = \frac{\sum_{k \in W_{out}^{-}} Sim_{out}(i,k)}{|W_{out}^{-}|}$$

where $W_{out}^{-} = \{k | s_{kj} = -1\}$ is the set of nodes that give negative edges to node j, and $Sim_{out}(i,k)$ is the similarity between node i and node k, calculated by Equation (15).

c) **Destination Node Positive Similarity** $S_{ji}^{+,in}$: The average similarity between node j and nodes that receive positive edges from i. The higher the $S_{ji}^{+,in}$ value, the greater the probability that node j receives positive edges from node i. $S_{ji}^{+,in}$ is defined as:

$$S_{ji}^{+,in} = \frac{\sum_{k \in W_{in}^{+}} Sim_{in}(j,k)}{|W_{in}^{+}|}$$

where $W_{in}^{+} = \{k | s_{ik} = 1\}$ is the set of nodes that receive positive edges from node i, and $Sim_{in}(j,k)$ is the similarity between node j and node k, calculated as:

$$Sim_{in}(j,k) = \frac{\sum_{p \in I_{jk}^{in}} r_{pj} \cdot r_{pk}}{\sqrt{\sum_{p \in I_{jk}^{in}} r_{pj}^2} \sqrt{\sum_{p \in I_{jk}^{in}} r_{pk}^2}}$$

where r_{pj} and r_{pk} are the edge signs from node p to nodes j and k, respectively, and I_{jk}^{in} is the set of nodes from which both j and k receive edges.

d) **Destination Node Negative Similarity** $S_{ji}^{-,in}$: The average similarity between node j and nodes that receive negative edges from i. The higher the $S_{ji}^{-,in}$ value, the greater the probability that node j receives negative edges from node i. $S_{ji}^{-,in}$ is defined as:

$$S_{ji}^{-,in} = \frac{\sum_{k \in W_{in}^{-}} Sim_{in}(j,k)}{|W_{in}^{-}|}$$

where $W_{in}^{-} = \{k | s_{ik} = -1\}$ is the set of nodes that receive negative edges from node i, and $Sim_{in}(j,k)$ is the similarity between node j and node k, calculated by Equation (18).

The logical relationships among the four node similarity quantification features are shown in Figure 2 [Figure 2: see original paper]. Based on logistic regression edge sign prediction, we establish the edge sign prediction model by fusing node status and similarity. Node status quantification features reflect nodes' social status and optimism degree from a global perspective, while node similarity quantification features reflect node attributes and preferences from a local perspective. By merging both as the feature set input, we achieve further model extension. The LR-SN model is defined as:

$$P(y=1|x;\theta)=\frac{1}{1+e^{-(\theta_0+\theta_1 Rep_i+\theta_2 Opt_i+\theta_3 Rep_j+\theta_4 Opt_j+\theta_5 S_{ij}^{+,out}+\theta_6 S_{ij}^{-,out}+\theta_7 S_{ji}^{+,in}+\theta_8 S_{ji}^{-,in})}}$$

4.4 Optimization Algorithm

We establish two models: the node status-based edge sign prediction model LR-S and the node status and similarity-based edge sign prediction model LR-SN. Node status quantifies sign attribute-related features from a global perspective, while node similarity embodies sign attributes from a local perspective. By merging both as the feature set input, we model the edge sign prediction problem using the logistic regression method in Equation (6). During model training, we use the stochastic gradient ascent algorithm for optimization. This algorithm updates the weight vector based on only one randomly selected sample per iteration, significantly reducing computational load and enabling faster training compared to traditional gradient ascent.

Combining Equations (7) and (8), we obtain the cost function for a single sample:

$$Cost(h_\theta(x),y)=-[y\log(h_\theta(x))+(1-y)\log(1-h_\theta(x))]$$

The cost function estimates the error between predicted and actual values. Given a sample, we can calculate the probability of that sample belonging to a certain class through the cost function. Assuming each sample is independent, the probability of the entire sample set is the product of all sample probabilities. For convenience, we take the logarithm to obtain the cost function for the entire sample set:

$$J(\theta)=\sum_{i=1}^{m}[y^{(i)}\log(h_\theta(x^{(i)}))+(1-y^{(i)})\log(1-h_\theta(x^{(i)}))]$$

where m is the total number of samples, $y^{(i)}$ represents the sign of the i-th sample, and $x^{(i)}$ represents the feature vector of the i-th sample.

To find the θ value that maximizes the cost function, we use the stochastic gradient ascent algorithm with the following update formula:

$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}$$

where j is the iteration count, α is the step size or learning rate controlling the update magnitude. Step size selection is crucial during training: if α is too large, it may cause divergence or non-convergence; if α is too small, it increases iteration count and slows convergence. To ensure stable algorithm progression, we reduce α by $\frac{1}{j+k}$ in each iteration, effectively solving problems such as oscillation near the optimal value caused by fixed step sizes. The specific steps are shown in Algorithm 1.

Algorithm 1: LR-SN 1. Input: Social network $G(V, E, S)$ 2. Input: Training sample set 3. while not converged do 4. Randomly select a sample $(x^{(i)}, y^{(i)})$ 5. Update weights: $\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}$ for $j \in \{0, 1, 2, ..., 8\}$ 6. Update step size: $\alpha = \frac{1}{j+k}$ 7. end while

---

5 Experiments

5.1 Dataset Description

We validate the effectiveness of the proposed models through experiments on three real-world datasets (Epinion, Slashdot, Wikipedia, available for download from the Snap website). Epinion is an online review website where people express opinions about each other using signs like 1 and -1. Slashdot is a technology news website where users can mark others as friends or foes. Wikipedia is a famous encyclopedia created by volunteers worldwide, where administrators are elected through voting and users can vote for or against candidates. Table 1 presents the statistical characteristics of the three real-world datasets.

Table 1: Dataset Statistics

Dataset	Positive Edges (%)	Negative Edges (%)	Average Clustering Coefficient
Epinion	78.7%	21.3%	0.22
Slashdot	85.4%	14.6%	0.06
Wikipedia	80.2%	19.8%	0.14

The statistical results in Table 1 show that the proportion of negative edges in all three networks is below 25%, with negative edges far outnumbered by positive edges. From psychology and sociology, people rarely express dislike or hatred toward other users in social networks due to politeness or fear of retaliation. Additionally, Epinion has the highest average clustering coefficient among the three networks, indicating its nodes are most densely distributed, followed by Wikipedia, while Slashdot has the lowest clustering coefficient.

5.2 Experimental Setup and Evaluation Metrics

As shown in the dataset statistics, the distribution of positive and negative edges in the three network datasets is extremely unbalanced, which may lead to low credibility in sign prediction accuracy. Therefore, we adopt random sampling during experiments to split datasets into training and testing sets, sequentially selecting 10%, 30%, 50%, ···, 90% of the dataset for training, with the remaining 90%, 70%, 50%, ···, 10% for testing. To ensure reliable prediction results, we repeat each experiment five times and report the average.

We use accuracy to evaluate the prediction algorithm' s performance on edge sign prediction, as shown in Equation (26). Additionally, we represent prediction results using a confusion matrix, as shown in Table 2 .

Table 2: Confusion Matrix

Real Value	Positive	Negative
Positive	TP (True Positive)	FN (False Negative)
Negative	FP (False Positive)	TN (True Negative)

5.3 Comparison of Edge Sign Prediction Models

To demonstrate the effectiveness of the proposed edge sign prediction models, we compare the LR-S and LR-SN models with three baseline methods:

a) **Status Method**: This method is defined based on the strong correlation between node status and edge sign attributes. It predicts signs according to the social status difference between two users. The greater the positive status gap, the more likely the edge sign is positive; the greater the negative status gap, the more likely the edge sign is negative.

b) **Balance Method**: This method is defined based on the strong correlation between structural balance theory [5] and edge sign attributes. During sign prediction, edge signs are inferred based on the structural balance of the triad containing the edge to be predicted. This theory relies on intuitive knowledge such as "a friend's friend is my friend"and "an enemy's friend is my enemy" [17], determining that a triangle is structurally balanced when it has an odd number of positive edges.

c) **LR Method [7]**: This method is proposed based on status theory and structural balance theory, extracting two types of features from the network. The first type includes degree features: the total out-degree d_i^{out} of node i, the positive out-degree d_i^{out+} and negative out-degree d_i^{out-} of node i, and the total in-degree d_j^{in} of node j, as well as the positive in-degree d_j^{in+} and negative in-degree d_j^{in-} of node j. The second type includes triad features, comprising 16 different triadic patterns where the edge to be predicted is located. These two types of features are trained using logistic regression to achieve sign prediction.

We apply the above methods for sign prediction on the Epinion, Slashdot, and Wikipedia datasets. The experimental comparison results are shown in Figure 3 [Figure 3: see original paper].

Figure 3: Comparison of Different Sign Prediction Methods

By analyzing Figures 3(a), 3(b), and 3(c), we can observe:

a) Both LR-S and LR-SN models achieve higher sign prediction accuracy than other baseline methods, significantly outperforming Status and Balance methods, demonstrating the effectiveness of node status and similarity quantification strategies. Moreover, the LR-SN model performs best in experiments, with prediction accuracy averaging 0.73% higher than LR-S and 3.32% higher than the LR method. This indicates that fusing local and global information through combining node status and similarity can effectively improve sign prediction accuracy.

b) Comparing LR-S and LR-SN models reveals that adding node similarity quantification features that consider local information improves prediction accuracy. For example, in the Epinion dataset, LR-SN achieves 96.31% prediction accuracy when the training set ratio is 90%, representing a 1.29% improvement over LR-S. The improvement is less pronounced in the other two datasets, likely because Slashdot and Wikipedia datasets are sparser, providing less local information.

c) Comparing Status and Balance methods shows that Balance outperforms Status in all three datasets, particularly in Epinion. This suggests that local information-based methods are more effective for sign attribute prediction than global information-based methods, especially in dense networks. Additionally, the LR method also achieves good prediction results by comprehensively considering status theory and structural balance theory, significantly improving accuracy over Status and Balance methods, further validating that fusing global and local information can enhance prediction accuracy.

Furthermore, we test the generalization ability of LR-S and LR-SN models. Both models trained on the three datasets demonstrate good generalization ability, as shown in Tables 3 and 4 .

Table 3: Generalization Ability of Sign Prediction Model LR-S (Training Ratio 90%)

Trained on	Epinion	Slashdot	Wikipedia
Epinion	95.02%	89.94%	88.63%
Slashdot	94.89%	90.12%	88.48%
Wikipedia	94.77%	89.65%	88.75%

Table 4: Generalization Ability of Sign Prediction Model LR-SN

(Training Ratio 90%)

Trained on	Epinion	Slashdot	Wikipedia
Epinion	96.31%	90.56%	88.96%
Slashdot	96.05%	90.81%	88.42%
Wikipedia	95.74%	90.29%	88.74%

5.4 Impact of Different Quantification Strategies on Edge Sign Prediction

To further validate the impact of node status and similarity on edge sign prediction models, we also use node similarity quantification features alone as the feature set input for sign prediction with logistic regression, denoted as LR-N. We then compare LR-S, LR-N, and LR-SN models. The comparison results are shown in Figure 4 [Figure 4: see original paper].

Figure 4: Comparison of Different Quantification Strategies

By analyzing Figure 4, we can draw the following conclusions:

a) The LR-SN model outperforms both LR-S and LR-N models across all three datasets. This validates on one hand that fusing local structural information and global structural information helps improve sign prediction accuracy, and on the other hand demonstrates the effectiveness of both node status and node similarity quantification strategies for sign prediction.

b) Both LR-S and LR-N models achieve high prediction performance in the three datasets, with LR-S performing better in Slashdot and Wikipedia datasets, while LR-N performs better in the Epinion dataset. This may be because among the three datasets, Epinion has the highest clustering coefficient and is the densest, providing more effective information for node similarity that considers local information. In contrast, Slashdot and Wikipedia datasets are sparser, making node status that considers global information more advantageous.

---

6 Conclusion

This paper explores the strong correlation between node status, similarity, and edge sign attributes, and implements edge sign prediction in social networks using the logistic regression method LR. We establish two models: the node status-based edge sign prediction model LR-S and the node status and similarity-based edge sign prediction model LR-SN. Node status quantifies sign attribute-related features from a global perspective, while node similarity embodies sign attributes from a local perspective. Experimental results on three real-world

network datasets demonstrate that the proposed models significantly improve sign prediction accuracy compared with existing baseline methods and exhibit certain generalizability. Future research will focus on further exploring factors influencing sign attributes and validating model performance using more real-world network datasets.

References

[1] Su Xiaoping, Song Yurong. Local signing features in signed networks and method of sign prediction [J]. Journal of Intelligent Systems, 2018, 13(3): 437-444.

[2] Tang Jiliang, Aggarwal C, Liu Huan. Recommendations in signed social networks [C]// Proc of the 25th International Conference on World Wide Web. 2016: 31-40.

[3] Li Dong, Xu Zhiming, Chakraborty N, et al. Polarity related influence maximization in signed social networks [J]. PLoS One, 2014, 9(7): e102199.

[4] Wang Peng, Song Yanhong, Li Songjiang, et al. Detection method for abnormal user of social network based on behavior characteristics [J]. Journal of Computer Applications, 2017, 37(S2): 219-224.

[5] Heider F. Attitudes and cognitive organization [J]. Journal of Psychology, 1946, 21(1): 107-112.

[6] Cartwright D, Harary F. Structural balance: a generalization of Heider' s theory [J]. Psychological Review, 1956, 63(5): 277-293.

[7] Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks [C]// Proc of the 19th International Conference on World Wide Web. New York: ACM Press, 2010: 641-650.

[8] Javari A, Mahdi J. Cluster-based collaborative filtering for sign prediction in social networks with positive and negative Links [J]. ACM Trans on Intelligent System and Technology, 2014, 5(2): 1-24.

[9] Symeonidis P, Tiakas E. Transitive node similarity: predicting and recommending links in signed social networks [J]. World Wide Web, 2014, 17(4): 743-776.

[10] Chiang K Y, Natarajan N, Tewari A, et al. Exploiting longer cycles for link prediction in signed network [C]//Proc of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2011: 1157-1162.

[11] Matsuo Y, Yamamoto H. Community gravity: measuring bidirectional effects by trust and rating on online social networks [C]// Proc of the 18th Int Conf on World Wide Web. New York: ACM Press, 2009: 641-650.

[12] Zolfaghar K, Aghaie A. Mining trust and distrust relationships in social Web applications [C]// Proc of the 6th International Conference on Intelligent Computer Communication and Processing. Piscataway, NJ: IEEE Press, 2010: 73-80.

[13] Shahriari M, Jalili M. Ranking nodes in signed social networks [J]. Social Network Analysis and Mining, 2014, 4(1): 1-12.

[14] Leskovec J, Huttenlocher D, Kleinberg J. Signed networks in social media[C]//Proc of SIGCHI Conference on Human Factors Computing Systems, New York: ACM Press, 2010: 1361-1370.

[15] Cheng Suqi, Shen Huawei, Zhang Guoqing, et al. Survey of signed networks research [J]. Journal of Software, 2014, 25(1): 1-15.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv —Machine translation. Verify with original.