2015-04

A Naïve Bayes model based on overlapping groups for link prediction in online social networks

Symposium on Applied Computing, 30th, 2015, Salamanca.
http://www.producao.usp.br/handle/BDPI/49012

A Naïve Bayes model based on ovelapping groups for link prediction in online social networks

Jorge Valverde-Rebaza, Alan Valejo, Lilian Berton, Thiago de Paulo Faleiros and Alneu de Andrade Lopes
Department of Computer Science
ICMC, University of São Paulo
C. P. 668, CEP 13560-970, São Carlos, SP, Brazil
{jvalverr, alan, lilian, thiagopf, alneu}@icmc.usp.br

ABSTRACT

Link prediction in online social networks is useful in numerous applications, mainly for recommendation. Recently, different approaches have considered friendship groups information for increasing the link prediction accuracy. Nevertheless, these approaches do not consider the different roles that common neighbors may play in the different overlapping groups that they belong to. In this paper, we propose a new approach that uses overlapping groups structural information for building a naïve Bayes model. From this proposal, we show three different measures derived from the common neighbors. We perform experiments for both unsupervised and supervised link prediction strategies considering the link imbalance problem. We compare sixteen measures in four well-known online social networks: Flickr, LiveJournal, Orkut and Youtube. Results show that our proposals help to improve the link prediction accuracy.

Categories and Subject Descriptors

E.1 [**Data**]: Data structures—*Graphs and networks*; H.2.8 [**Database Applications**]: Data mining; H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Link Prediction, Social Networks, Overlapping Community, Naïve Bayes Model

1. INTRODUCTION

The boundless growth of online social networks has resulted in several research directions that examine structural and other properties of large-scale social networks. One of the most relevant research in social networks is the link prediction [4, 7, 10, 5]. Link prediction addresses the problem of predicting the likelihood of existence of future links between disconnected nodes. Several methods have been proposed to cope with this problem. Most of them assign a score for each pair of nodes based on its neighborhood nodes (local) or path (global) information [4, 7].

The performance of local measures, such as Common Neighbors (CN), Adamic Adar (AA), Jaccard Coefficient (Jac), Resource Allocation (RA) and Preferential Attachment (PA) was compared to the performance of global measures, such as Katz, simRank and rooted PageRank, in [4] and [7]. According to these experimental results on real networks, global measures usually achieve higher accuracy than the local ones. Nevertheless, global measures are very time-consuming and usually infeasible for large-scale networks.

Recently, aiming to improve the link prediction accuracy, different methods have been proposed considering measures based on the naïve Bayes model and on community information. Measures based on the naïve Bayes model [6], such as Local Naïve Bayes (LNB) and their Common Neighbors, Adamic Adar, and Resource Allocation forms (LNB-CN, LNB-AA, and LNB-RA, respectively) capture different roles of common neighbors and assign to them different weights.

Measures based on community information capture the correlation between nodes belonging to the same communities [10, 15, 9]. Most of these methods consider that a node belongs to just one community. However, in online social networks users usually belong to more than one community. Thus, in [12] are proposed three measures based on overlapping communities information: Common Neighbors Within and Outside of Common Groups (WOCG), Common Neighbors of Groups (CNG) and Common Neighbors with Total and Partial Overlapping of Groups (TPOG). The assumption of these measures is that common neighbors which belong to the same group of a pair of vertices are likely more influential in the existence of a future link between these vertices than common neighbors of different groups.

In this paper, we analyze links and overlapping community structure in four large-scale online social networks to deal with the link prediction problem. Our contributions are three fold: 1) From the overlapping community structure, we propose two new network features: the overlapping groups degree and the overlapping groups clustering coefficient. 2) Based on these new network features and using a naïve Bayes model, we propose a new link prediction measure, the Group Naïve Bayes measure (GNB) and its three forms derived from Common Neighbors (GNB-CN), Adamic Adar (GNB-AA) and Resource Allocation (GNB-RA). 3) We compared the performance of our proposals with other supervised and unsupervised techniques described in the literature.

SAC'15 April 13-17, 2015, Salamanca, Spain.

http://dx.doi.org/10.1145/2695664.2695719

The remainder of this paper is organized as follows. In Section 2, we present the description of the link prediction problem. In Section 3, we present and explain our proposals. In Section 4, we present experimental results obtained from four online social networks. Finally, in Section 5, we summarize the main findings and conclusions of this work.

2. PROBLEM DESCRIPTION

The link prediction problem can be approached by unsupervised and supervised strategies.

2.1 Unsupervised Strategy

Given a network $G = (V, E)$, where V and E are sets of nodes and links respectively. Multiple links and self-connections are not allowed. If G is a directed network, consider the universal set, denoted by U, containing all $|V|(|V|-1)$ potential directed links between pair of nodes in V, where $|V|$ denotes the number of elements in V. If G is an undirected network, the universal set U contains $\frac{|V|(|V|-1)}{2}$ links. The fundamental link prediction task in the unsupervised context is to find out the missing links (future links) in the set $U - E$ (set of nonexistent links) assigning a score for each link in this set. The higher the score, the higher the connection probability, and vice versa [7, 14, 10].

Two standard evaluation measures are used to quantify the prediction accuracy [7, 11, 12]: AUC (area under the receiver operating characteristic curve) and precision. The AUC is interpreted as the probability that a randomly chosen and correctly predicted link has a higher score than other randomly chosen and wrongly predicted link. Thus, for n independent comparisons, if n' times for the links correctly predicted are given higher scores than for links wrongly predicted whilst n'' times for both correctly and wrongly predicted links are given equal scores, the AUC is defined as $AUC = \frac{n'+0.5n''}{n}$.

Different from AUC, precision only focuses on the L links with highest scores. Thus, precision is defined as the ratio between the L_r correctly predicted links from the L top-ranked links, i.e. $precision = \frac{L_r}{L}$.

2.2 Supervised Strategy

Supervised strategy considers the link prediction problem as a classification problem. Thus, network information such as the structural ones and nodes attributes are used to build a set of feature vectors for both linked and not linked pairs of nodes [3, 1, 5, 11].

By using the supervised strategy is possible to employ different validation processes, such as k-fold cross-validation [3]. Thus, we can use the traditional evaluation measures, such as accuracy, precision, recall, f-measure and AUC, to compare classifiers performances [2].

It is important to notice that evaluation measures for unsupervised strategy, such as the precision and AUC, are calculated differently than for supervised strategy but in both cases they have essentially the same meaning. Furthermore, unsupervised evaluation measures are applied directly on results of link prediction measures but supervised evaluation measures are applied on results of classifiers [11].

3. OVERLAPPING GROUP INFORMATION FOR LINK PREDICTION

In this section we introduce some new concepts related to overlapping groups on the network structure. After that, we present our four link prediction proposals.

3.1 Preliminary

Given the network G with $M > 1$ groups identified by different group labels $g_1, g_2, \ldots, g_M$. Each node $x \in V$ belongs to a set of node groups $\mathcal{G}_\alpha = \{g_a, g_b, \ldots, g_p\}$ with size P. Thus, $P > 0$ and $P \leq M$. Each $g_i \in \mathcal{G}$ is a group of nodes, whose elements share interests and behaviors. With M groups in G is possible to form N different sets of groups $\mathcal{G}_\alpha, \mathcal{G}_\beta, \ldots, \mathcal{G}_N$. When a node x belongs to a set of node groups $\mathcal{G}_\alpha$, this node is represented as $x^{\mathcal{G}_\alpha}$. A node belongs to just one set of node groups [12].

The basic structural definition for a node $x \in V$ is its neighborhood $\Gamma(x) = \{y \mid (x, y) \in E \vee (y, x) \in E\}$ which denotes the set of neighbors of x. For a pair of nodes (x, y), $\Lambda_{x,y} = \Gamma(x) \cap \Gamma(y)$ denotes its set of common neighbors [4, 10, 12].

We define the *overlapping groups neighborhood* of a node $x^{\mathcal{G}_\alpha}$, $\Gamma^{\mathcal{G}}(x) = \{y^{\mathcal{G}_\beta} \mid ((x^{\mathcal{G}_\alpha}, y^{\mathcal{G}_\beta}) \in E \vee (y^{\mathcal{G}_\beta}, x^{\mathcal{G}_\alpha}) \in E) \wedge \mathcal{G}_\alpha \cap \mathcal{G}_\beta \neq \varnothing\}$, as the set of neighbors of x belonging to some groups to which x belongs to. For a pair of disconnected nodes $(x^{\mathcal{G}_\alpha}, y^{\mathcal{G}_\beta})$, denote by $\Lambda^{\mathcal{G}}_{x,y} = \Gamma^{\mathcal{G}}(x) \cap \Gamma^{\mathcal{G}}(y)$ its set of common neighbors of groups.

The cardinality of overlapping groups neighborhood set defines the *overlapping groups degree of a vertex* x , which is denoted by $k^{\mathcal{G}}(x) = |\Gamma^{\mathcal{G}}(x)|$. The average of the overlapping groups degree of all nodes in G is called as the *average overlapping groups degree*, $\langle k^{\mathcal{G}} \rangle$.

Considering the above definitions and formalisms showed by [8] and [6], we define the *overlapping groups clustering coefficient* of a node x as the clustering coefficient of the subgraph consisting only of nodes belonging to its overlapping groups neighborhood. Thus, the overlapping groups clustering coefficient of x can be calculated by:

$$C^{\mathcal{G}}_x = \frac{\Delta^{\mathcal{G}}_x}{\Delta^{\mathcal{G}}_x + \Lambda^{\mathcal{G}}_x} \quad (1)$$

where $\Delta^{\mathcal{G}}_x$ and $\Lambda^{\mathcal{G}}_x$ are respectively the number of connected and disconnected pair of nodes whose common neighbors of groups include x. Clearly, $\Delta^{\mathcal{G}}_x + \Lambda^{\mathcal{G}}_x = \frac{k^{\mathcal{G}}(x)(k^{\mathcal{G}}(x)-1)}{2}$. The *average overlapping groups clustering coefficient*, $C^{\mathcal{G}}$, is the average of the overlapping groups clustering coefficient of all nodes in G.

3.2 Method

Following the formalism showed in [12], for a network G, we denote by $L_{x,y}$ and $\overline{L}_{x,y}$ the class variables of link existence and nonexistence, respectively, for a pair of nodes $(x, y) \in V$. The prior probabilities of $L_{x,y}$ and $\overline{L}_{x,y}$ are calculated according to Eq. 2 and 3, respectively.

$$P(L_{x,y}) = \frac{|E|}{|U|} \quad (2) \qquad P(\overline{L}_{x,y}) = \frac{|U| - |E|}{|U|} \quad (3)$$

Each node z owns two conditional probabilities, $P(z \mid L_{x,y})$, which is the probability that node z is the common neighbor of groups of a connected pair (x, y), and $P(z \mid \overline{L}_{x,y})$ is the probability that node z is the common neighbor of groups of a disconnected pair (x, y). According to Bayesian theory,

these two probabilities are:

$$P(z \mid L_{x,y}) = \frac{P(z)P(L_{x,y} \mid z)}{P(L_{x,y})} \quad (4)$$

$$P(z \mid \overline{L}_{x,y}) = \frac{P(z)P(\overline{L}_{x,y} \mid z)}{P(\overline{L}_{x,y})} \quad (5)$$

The posterior probability of connection and disconnection of the pair (x, y) given its set of common neighbors of groups are:

$$P(L_{x,y} \mid \Lambda^{\mathcal{G}}_{x,y}) = \frac{P(L_{x,y})P(\Lambda^{\mathcal{G}}_{x,y} \mid L_{x,y})}{P(\Lambda^{\mathcal{G}}_{x,y})} \quad (6)$$

$$P(\overline{L}_{x,y} \mid \Lambda^{\mathcal{G}}_{x,y}) = \frac{P(\overline{L}_{x,y})P(\Lambda^{\mathcal{G}}_{x,y} \mid \overline{L}_{x,y})}{P(\Lambda^{\mathcal{G}}_{x,y})} \quad (7)$$

In order to compare the existence likelihood between node pairs, we define the likelihood score, $s_{x,y}$, of a node pair (x, y) as the ratio between Eq. 6 and 7. Based on the naïve Bayes scheme showed in [6], we decompose $P(\Lambda^{\mathcal{G}}_{x,y} \mid L_{x,y}) = \prod_{z \in \Lambda^{\mathcal{G}}_{x,y}} P(z \mid L_{x,y})$ and $P(\Lambda^{\mathcal{G}}_{x,y} \mid \overline{L}_{x,y}) = \prod_{z \in \Lambda^{\mathcal{G}}_{x,y}} P(z \mid \overline{L}_{x,y})$. Thus, substituting Eqs. 4 and 5, we have:

$$s_{x,y} = \frac{P(L_{x,y})}{P(\overline{L}_{x,y})} \prod_{z \in \Lambda^{\mathcal{G}}_{x,y}} \frac{P(\overline{L}_{x,y})P(L_{x,y} \mid z)}{P(L_{x,y})P(\overline{L}_{x,y} \mid z)} \quad (8)$$

Indeed $P(L_{x,y} \mid z)$ is equal to the overlapping groups clustering coefficient of node z, as stated in Eq. 9. Since $P(L_{x,y} \mid z) + P(\overline{L}_{x,y} \mid z) = 1$, using the Eq. 1, $P(\overline{L}_{x,y} \mid z)$ is calculated as stated in Eq. 10.

$$P(L_{x,y} \mid z) = C^{\mathcal{G}}_z \quad (9)$$

$$P(\overline{L}_{x,y} \mid z) = 1 - C^{\mathcal{G}}_z = \frac{\Lambda^{\mathcal{G}}_z}{\Delta^{\mathcal{G}}_z + \Lambda^{\mathcal{G}}_z} \quad (10)$$

Substituting Eqs. 2, 3, 9 and 10 into Eq. 8, the likelihood score of a node pair (x, y) is:

$$s_{x,y} = \Omega \prod_{z \in \Lambda^{\mathcal{G}}_{x,y}} \Omega^{-1} \frac{\Delta^{\mathcal{G}}_z}{\Lambda^{\mathcal{G}}_z} \quad (11)$$

where $\Omega = \frac{P(L_{x,y})}{P(\overline{L}_{x,y})} = \frac{|E|}{|U|-|E|}$ is a constant for a network and its computation can be disregarded. To prevent the division by zero, we can use any smoothing method. Thus, using the add-one smoothing, we define the **group naïve Bayes** (GNB) measure as:

$$s^{GNB}_{x,y} = \prod_{z \in \Lambda^{\mathcal{G}}_{x,y}} \Omega^{-1} N^{\mathcal{G}}_z \quad (12)$$

where $N^{\mathcal{G}}_z = \frac{\Delta^{\mathcal{G}}_z + 1}{\Lambda^{\mathcal{G}}_z + 1}$. Clearly, larger score means higher probability that two nodes are connected.

3.3 Group Naïve Bayes Forms

The connection likelihood between a pair of nodes can be improved by identifying the different roles that their common neighbors play, for example, identifying their behaviors into the different groups that they belong to [10, 12]. Hence, measures such as CN, AA and RA try to capture different roles from the set of all common neighbors. Thus, we adapt the GNB measure to capture these roles too.

Following the formalism showed in [6], we add an exponent $f(k^{\mathcal{G}}(x))$ to $\Omega^{-1} N^{\mathcal{G}}_z$ in Eq. 12, where f is a function of overlapping groups degree. Using Log function on both sides, we obtain the next linear equation:

$$s^{GNB'}_{x,y} = \sum_{z \in \Lambda^{\mathcal{G}}_{x,y}} f(k^{\mathcal{G}}(z)) \log(\Omega^{-1} N^{\mathcal{G}}_z) \quad (13)$$

Here we consider three forms of function f: $f(k^{\mathcal{G}}(x)) = 1$, $f(k^{\mathcal{G}}(x)) = \frac{1}{\log(k^{\mathcal{G}}(x))}$ and $f(k^{\mathcal{G}}(x)) = \frac{1}{k^{\mathcal{G}}(x)}$, which are corresponding to the group naïve Bayes form of CN, AA and RA, respectively:

$$s^{GNB-CN}_{x,y} = |\Lambda^{\mathcal{G}}_{x,y}| \log(\Omega^{-1}) + \sum_{z \in \Lambda^{\mathcal{G}}_{x,y}} \log(N^{\mathcal{G}}_z) \quad (14)$$

$$s^{GNB-AA}_{x,y} = \sum_{z \in \Lambda^{\mathcal{G}}_{x,y}} \frac{1}{\log(k^{\mathcal{G}}(z))} (\log(N^{\mathcal{G}}_z) + \log(\Omega^{-1})) \quad (15)$$

$$s^{GNB-RA}_{x,y} = \sum_{z \in \Lambda^{\mathcal{G}}_{x,y}} \frac{1}{k^{\mathcal{G}}(z)} (\log(N^{\mathcal{G}}_z) + \log(\Omega^{-1})) \quad (16)$$

4. EXPERIMENTS

We consider a scenario where new links of four online social networks must be predicted. We use the natural information provided by users that participate freely in different user groups to assign group labels to each node. We also compare the performance of our proposals to other measures based on local information, overlapping groups information and naïve Bayes model.

4.1 Datasets

Social networks considered in our experiments are Flickr, LiveJournal, Orkut and Youtube and are available in [8]. These networks have information of links between users and natural information about friendship groups to which each user belongs.

High-level topological features of the four social networks are presented in Table 1 [8]. From this table, we observe that due to the high number of nodes ($|V|$) and links ($|E|$) these networks are considered as large-scale networks. The average degree ($\langle k \rangle$) indicates the average of number of neighbors per user. The fraction of links symmetric ($\mathcal{S}$) denotes the degree in which directed links from a source to a destination have an endorsement of the destination by the source.

Table 1: Topological features of social networks

	Flickr	**LiveJournal**	**Orkut**	**Youtube**
$\|V\|$	1,846,198	5,284,457	3,072,441	1,157,827
$\|E\|$	22,613,981	77,402,652	223,534,301	4,945,382
$\langle k \rangle$	12.24	16.97	106.1	4.29
$\mathcal{S}$	62.0%	73.5%	100.0%	79.1%
$\langle l \rangle$	5.67	5.88	4.25	5.10
D	27	20	9	21
C	0.313	0.330	0.171	0.136
r	0.202	0.179	0.072	−0.033
M	103,648	7,489,073	8,730,859	30,087
$\langle m \rangle$	4.62	21.25	106.44	0.25
$\langle P \rangle$	82	15	37	10
$\langle g \rangle$	0.47	0.81	0.52	0.34
$\langle k^{\mathcal{G}} \rangle$	9.65	6.19	50.85	0.42
$C^{\mathcal{G}}$	0.06	0.13	0.18	0.02

Also, Table 1 shows global topological features of networks. The average path length ($\langle l \rangle$) is the average number

of steps along the shortest paths for all possible node pairs and the diameter (D) is defined as the maximum shortest path between any two nodes. The average clustering coefficient (C) is the degree to which nodes in a network tend to cluster together and the assortativity coefficient (r) indicates the likelihood for nodes to connect to other nodes with similar degrees.

Among the group features, we observe that the four networks have a high amount of groups (M) and that users, in LiveJournal and Orkut, belong to a high number of groups ($\langle m \rangle$) (except Youtube and Flickr). The average group size ($\langle P \rangle$) is at least of 10 users. The average group clustering coefficient ($\langle g \rangle$), is defined as the average of clustering coefficients of the subgraphs consisting of only the users who are members of each group. The average overlapping groups degree $\langle k^{\mathcal{G}} \rangle$ and the average overlapping groups clustering coefficient ($C^{\mathcal{G}}$) were defined in Section 3.1.

4.2 Experimental Setup

For the network preprocessing, for a network G, the set E is divided into the training set E^T and the probe set E^P. From the set E, for selecting the links for E^P, we take randomly two-third of the links formed by nodes whose number of neighbors is two times greater than the average degree per node. The remaining links, except those formed by nodes whose number of neighbors is less than two-third of the average degree per node, constitute the training set E^T. This evaluation method is widely used in the link prediction literature [11, 12, 13, 14].

After that, the link prediction process is initiated. This process includes both unsupervised and supervised strategies. In unsupervised strategy, for each pair of nodes from E^T, the connection likelihood is calculated based on the link direction, choosing the highest score between its *in* and *out* scores as final and unique score, e.g., by vertex pair (x, y) if $s^{out}_{x,y} > s^{in}_{x,y}$ then $s_{x,y} = s^{out}_{x,y}$, otherwise, $s_{x,y} = s^{in}_{x,y}$ [12, 14].

In supervised strategy, we use decision tree (J48), naïve Bayes (NB), multilayer perceptron with backpropagation (MLP) and support vector machine (SMO) classifiers from Weka[1]. Thus, for each network, we compute a set of feature vector formed by randomly selected pair of nodes from E^T. If the pair of nodes taken from the predicted links list from E^T is also in E^P then the feature vector formed by this pair of nodes takes the positive class (existent link), otherwise takes the negative class (nonexistent link). Table 2 shows the number of instances by class and the total of instances for each social network. Note that we consider an imbalanced class distribution.

Table 2: Number of instances by class

	Existent	Non-existent	Total
Flickr	7,100	35,500	42,600
LiveJournal	4,500	22,500	27,000
Orkut	16,000	80,000	96,000
Youtube	2,700	13,500	16,200

For each network, we create ten different data sets. Each data set is formed by features which combine different link prediction measures as specified in Table 3.

4.3 Results

We perform experiments to validate the link prediction in both unsupervised and supervised context. For both cases we apply the evaluation measures presented in Section 2 on sixteen link prediction measures, which are grouped in four sets: i) local measures: CN, RA, AA, PA and Jac; ii) measures based on overlapping groups information [12]: WOCG, CNG and TPOG; iii) measures based on the local naïve Bayes model [6]: LNB, LNB-CN, LNB-AA and LNB-RA; and iv) measures based on the overlapping groups naïve Bayes model: GNB, GNB-CN, GNB-AA and GNB-RA.

Table 3: Data sets created for each network

Data set	Features
VLocal	CN, AA, Jac, RA and PA
VGroups	WOCG, CNG and TPOG
VLNB	LNB, LNB-CN, LNB-AA and LNB-RA
VGNB	GNB, GNB-CN, GNB-AA and GNB-RA
VLocal-Groups	VLocal and VGroups
VLocal-GNB	VLocal and VGNB
VLNB-Groups	VLNB and VGroups
VLNB-GNB	VLNB and VGNB
VGroups-GNB	VGroups and VGNB
VTotal	VLocal, VGroups, VLNB and VGNB

4.3.1 Unsupervised evaluation

AUC and precision were employed for analysing results of unsupervised link prediction process. Table 4 summarizes the prediction results measured by AUC, with $n = 5000$. Each AUC value is obtained by averaging over 10 run over 10 independent partitions of training and testing sets. For each network, values highlighted in gray indicate the highest result for each type of measure and values emphasized in bold correspond to the highest AUC achieved. The last column shows the average ranking of each measure.

Table 4: The prediction results measured by AUC

Method	Flickr	Livejournal	Orkut	Youtube	Avg. rank
CN	0.674	0.582	0.572	0.834	10.50
AA	0.656	0.580	0.620	**0.928**	8.25
Jac	0.431	0.624	0.575	0.217	12.50
RA	0.616	0.565	0.566	0.892	11.00
PA	0.566	0.542	0.602	0.917	10.00
WOCG	0.637	0.596	0.649	0.434	10.75
CNG	0.728	0.611	0.621	0.723	9.63
TPOG	0.728	0.665	0.651	0.555	8.63
LNB	0.860	0.880	0.446	0.872	7.25
LNB-CN	0.859	0.877	**0.706**	0.873	4.50
LNB-AA	0.884	**0.883**	0.342	0.890	5.75
LNB-RA	**0.890**	0.880	0.333	0.896	5.75
GNB	0.857	0.853	0.525	0.800	10.0
GNB-CN	0.861	0.855	0.639	0.808	6.25
GNB-AA	0.875	0.862	0.572	0.807	6.75
GNB-RA	0.874	0.856	0.539	0.790	8.50

Considering the performance of each type of measure, among the local measures, AA performs better. Among the measures based on overlapping groups information, TPOG performs better. Among the measures based on the local naïve Bayes model, LNB-RA performs better. Among the measures based on the overlapping groups naïve Bayes model, GNB-CN and GNB-AA perform better. Considering the best AUC for each network, for Flickr, LiveJournal and Orkut, the measures based on the local naïve Bayes model outperform the others. For Youtube, AA performs better.

Based on results of Table 4, Friedman and Nemenyi posthoc tests were applied to analyze the difference between all link prediction measures evaluated. The Friedman test using the F-statistics indicated the null-hypothesis, that all link prediction measures evaluated behave similarly, should not be rejected. So, there is no significant difference. On the

[1]http://www.cs.waikato.ac.nz/ml/weka/

top of the presented diagram is the critical difference (CD) value and in the axis are the average rank of measures. The lowest (best) ranks are in the left side of the axis.

Figure 1 presents the Nemenyi test for all sixteen measures. The critical value of the F-statistics with 15 and 45 degrees of freedom at 95 percentile is 1.89. According to the Nemenyi statistics, the CD for comparing the average ranking of two different link prediction measures at 95 percentile is 11.53. All measures analyzed have no significant difference, so they are connected by a bold line in the diagram.

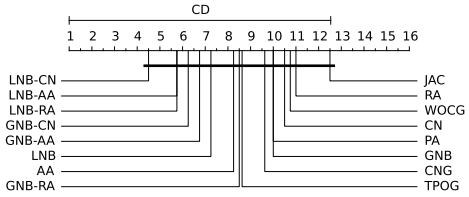

Figure 1: Post-hoc test for results from Table 4

Although there is no significant difference among them, we achieved a competitive accuracy with literature. Thus, in general terms, the measures based on the local naïve Bayes model and on the overlapping groups naïve Bayes model, specifically LNB-CN and GNB-CN, were first ranked. Following them, the measures based on overlapping groups information, specifically TPOG, and finally the local measures, specifically AA.

Due to the fact that measures based on local and overlapping groups naïve Bayes models achieve the best scores in the analysis of AUC in Figure 2 we show the precision results only of these measures. Different values of L are used. For Flickr, all link prediction measures have a similar performance, with maximum precision value equal to 0.4. For LiveJournal, LNB and GNB have the best overall measures in all L values, with maximum value equal to 0.8. For Orkut, LNB-CN, GNB-CN and GNB-AA have a similar precision performance with maximum value equal to 0.7. They reach their maximum performance for all L values. For Youtube, all our proposals, GNB, GNB-CN, GNB-AA and GNB-RA, achieve the best precision performance, with maximum value equal to 0.7. Most of the measures reach their maximum performance for $L = 100$ with a declining performance after this L value.

From our unsupervised evaluation, we observe that measures based on the local naïve Bayes model perform better on networks with high C and r values, such as Flickr and Livejournal. High values of $C^{\mathcal{G}}$ and $\langle k^{\mathcal{G}} \rangle$, such as in LiveJournal and Orkut, lead to a good performance of measures based on the overlapping groups naïve Bayes model. Youtube is a disassortative network (negative value of r) but with a high value of $\langle k^{\mathcal{G}} \rangle$ considering its small value of $\langle m \rangle$, so measures based on the overlapping groups naïve Bayes model have a good performance too.

4.3.2 *Supervised evaluation*

Due to the presence of an imbalanced class distribution in the data sets summarized in Table 3, we employ the AUC for analyze the results of supervised link prediction process. The average values for AUC considering a 10-fold cross validation process for J48, NB, MLP and SMO classifiers are shown in Table 5. Values emphasized in bold correspond to

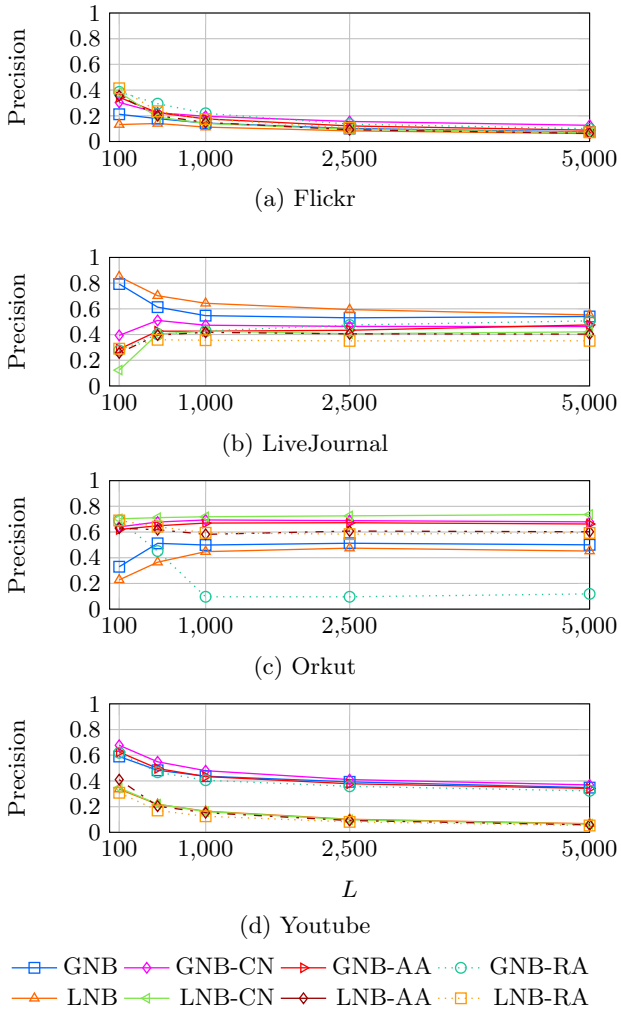

Figure 2: Precision results on four social networks. Different values of L are used to select the top-L highest scores for predicting links.

the highest results among the evaluated data sets for each classifier. Values highlighted in gray indicate that a classifier get best results in data sets using overlapping group information than VLocal data set.

From Table 5 one can observe that in most cases the best AUC is obtained by VLocal-Groups and VTotal. Also, that in any network neither VLNB nor VGNB were able to overcome VLocal but in several cases data sets formed by measures using overlapping groups information combined with local measures, i.e. VLocal-Groups and VLocal-GNB, outperform VLocal.

In order to observe the impact of the link prediction measures on the performance of classifiers, in Figure 3 is shown the Nemenyi test for all data sets of Flickr network. The critical value of the F-statistics with 9 and 27 degrees of freedom at 95 percentile is 2.25. According to the Nemenyi statistics, the CD for comparing the mean-ranking of two different link prediction measures at 95 percentile is 6.77. The measures that have no significant difference are connected by a bold line in the diagram. VTotal is better ranked followed by VLocal-Groups and VLocal-GNB, in second and third place, respectively. The main observation is the signi-

Table 5: Classifiers results measured by AUC

Network	Data set	J48	NB	SMO	MLP
Flickr	VLocal	0.774	0.746	0.583	0.778
	VGroups	0.761	0.728	0.504	0.734
	VLNB	0.748	0.664	0.501	0.685
	VGNB	0.737	0.502	0.501	0.516
	VLocal-Groups	0.789	**0.776**	0.585	0.778
	VLocal-GNB	**0.796**	0.725	0.583	0.780
	VLNB-Groups	0.792	0.723	0.504	0.753
	VLNB-GNB	0.769	0.642	0.502	0.688
	VGroups-GNB	**0.796**	0.698	0.505	0.736
	VTotal	0.793	0.747	**0.586**	**0.782**
Livejournal	VLocal	**0.808**	**0.829**	0.658	0.854
	VGroups	0.767	0.768	0.607	0.777
	VLNB	0.732	0.776	0.547	0.800
	VGNB	0.775	0.503	0.503	0.510
	VLocal-Groups	0.802	0.826	0.654	0.854
	VLocal-GNB	0.807	0.828	0.660	0.852
	VLNB-Groups	0.783	0.806	0.612	0.835
	VLNB-GNB	0.804	0.767	0.550	0.798
	VGroups-GNB	0.768	0.772	0.609	0.781
	VTotal	0.799	0.825	**0.664**	**0.858**
Orkut	VLocal	**0.883**	0.862	0.629	0.873
	VGroups	0.829	0.870	0.626	0.863
	VLNB	0.823	0.837	0.558	0.859
	VGNB	0.816	0.500	0.500	0.532
	VLocal-Groups	0.880	0.872	**0.644**	0.871
	VLocal-GNB	0.857	0.862	0.629	**0.876**
	VLNB-Groups	0.872	0.869	0.634	0.861
	VLNB-GNB	0.828	0.830	0.558	0.858
	VGroups-GNB	0.830	0.856	0.626	0.863
	VTotal	0.861	**0.873**	**0.644**	0.873
Youtube	VLocal	0.836	0.801	0.551	0.808
	VGroups	0.734	0.671	0.562	0.726
	VLNB	0.832	0.687	0.507	0.739
	VGNB	0.802	0.506	0.501	0.499
	VLocal-Groups	0.822	**0.819**	**0.579**	**0.825**
	VLocal-GNB	**0.851**	0.800	0.551	0.812
	VLNB-Groups	0.822	0.720	0.562	0.755
	VLNB-GNB	0.835	0.683	0.509	0.738
	VGroups-GNB	0.820	0.681	0.562	0.723
	VTotal	0.823	0.768	0.578	0.821

ficative differences between VTotal with VLNB and VGNB and between VLocal-Groups and VLocal-GNB with VGNB.

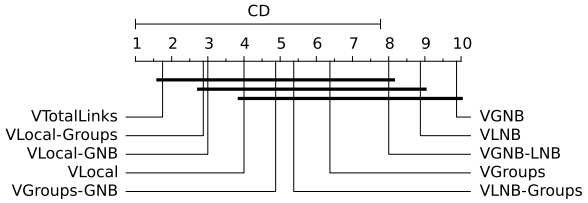

Figure 3: Post-hoc test for results from Table 5 for Flickr network

5. CONCLUSIONS

In this paper, based on a naïve Bayes model, four new link prediction measures were proposed considering the actual scenario of online social networks where users participate in overlapping groups

From the experiments performed we observe that, individually the local Naïve Bayes model and the overlapping groups Naïve Bayes model measures outperform those based only on overlapping group information and local information. Moreover, when local measures are combined with measures based on overlapping groups and on overlapping groups Naïve Bayes model, the link prediction accuracy improves.

Thus, our results suggest that using overlapping groups information improves the link prediction accuracy. Furthermore, we showed a comparison among the most usual measures, which serves as a guide to future researchers.

Acknowledgments

This work was partially supported by the São Paulo Research Foundation (FAPESP) grants: 2013/12191 − 5, 2011/21880 − 3, 2011/23689 − 9 and 2011/22749 − 8.

6. REFERENCES

[1] N. Benchettara, R. Kanawati, and C. Rouveirol. A supervised machine learning link prediction approach for academic collaboration recommendation. In *RecSys '10*, pages 253–256, 2010.

[2] M. Fatourechi, R. K. Ward, S. G. Mason, J. Huggins, A. Schlogl, and G. E. Birch. Comparison of evaluation metrics in classification applications with imbalanced datasets. In *ICMLA '08*, pages 777–782, 2008.

[3] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM '06 Workshop on Link Analysis, Counterterrorism and Security*, 2006.

[4] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.

[5] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *ACM SIGKDD KDD '10*, pages 243–252. ACM, 2010.

[6] Z. Liu, Q.-M. Zhang, L. Lü, and T. Zhou. Link prediction in complex networks: A local naïve bayes model. *EPL*, 96(4):48007, 2011.

[7] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150 – 1170, 2011.

[8] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *ACM SIGCOMM IMC '07*, pages 29–42. ACM, 2007.

[9] S. Soundarajan and J. Hopcroft. Using community information to improve the precision of link prediction methods. In *WWW '12*, pages 607–608. ACM, 2012.

[10] J. Valverde-Rebaza and A. Lopes. Link prediction in complex networks based on cluster information. In *SBIA '12*, pages 92–101. Springer-Verlag, 2012.

[11] J. Valverde-Rebaza and A. Lopes. Exploiting behaviors of communities of Twitter users for link prediction. *SNAM*, 3(4):1063–1074, 2013.

[12] J. Valverde-Rebaza and A. Lopes. Link prediction in online social networks using group information. In *ICCSA 2014*, volume 8584, pages 31–45, 2014.

[13] D. Yin, L. Hong, and B. D. Davison. Structural link analysis and prediction in microblogs. In *CIKM '11*, pages 1163–1168, 2011.

[14] Q.-M. Zhang, L. Lü, W.-Q. Wang, Y.-X. Zhu, and T. Zhou. Potential theory for directed networks. *PLoS ONE*, 8(2):e55437, 2013.

[15] E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter. Using friendship ties and family circles for link prediction. In *SNAKDD '08*, pages 97–113, 2008.