Journal of Multivariate Analysis 184 (2021) 104732

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Depth-based classification for relational data with multiple attributes

Xu Zhang [a,b,1], Yahui Tian [c], Guoyu Guan [d,2], Yulia R. Gel [e,*,3]

[a] School of Mathematical Sciences, South China Normal University, Guangzhou, China
[b] Key Laboratory for Applied Statistics of the MOE, and School of Mathematics and Statistics, Northeast Normal University, Changchun, China
[c] Department of Biostatistics and Data Sciences, Boehringer Ingelheim, China
[d] School of Economics and Management, Northeast Normal University, Changchun, China
[e] Department of Mathematical Sciences, University of Texas in Dallas, USA

ARTICLE INFO

Article history:
Received 24 July 2020
Received in revised form 27 January 2021
Accepted 27 January 2021
Available online 11 February 2021

AMS 2010 subject classifications:
62G99
62H30
62P25

Keywords:
Classification
Complex network
Data depth
Nonparametrics
Relational data

ABSTRACT

With the recent progress of data acquisition technology, classification of data exhibiting relational dependence, from online social interactions to multi-omics studies to linkage of electronic health records, continues to gain an ever increasing attention. By introducing a robust and inherently geometric concept of data depth we propose a new type of geometrically-enhanced classification method for relational data that are in a form of a complex network with multiple node attributes. Starting from a logistic regression to describe the relationship between the class labels and node attributes, the key approach is based on modeling the link probability between any two nodes as a function of their class labels and their data depths within the respective classes. The approximate prediction rule is then obtained according to the posterior probability of the class labels. Integrating the depth concept into the classification process allows us to better capture the underlying geometry of the relational data and, as a result, to enhance its finite sample performance. We derive asymptotic properties of the new classification approach and validate its finite sample properties via extensive simulations. The proposed geometrically-enhanced classification method is illustrated in application to user analysis of the one of the largest Chinese social media platforms, Sina Weibo.

1. Introduction

Classification constitutes one of the key tasks in modern statistical and data sciences, with methods ranging from more conventional tools such as logistic regression and linear discriminant analysis, to advanced machine learning techniques such as deep learning, see, e.g., [13,14,22,27,28]. Many traditional classification methods tend to rely on the assumption of independence among subjects. However, in the real world, "connections" or "relationships" are typically inevitable, and information conveyed by the link often provides an important insight into the classification process. With the recent

* Correspondence to: 800 West Campbell Road, Richardson, TX 75080, USA.
E-mail address: ygl@utdallas.edu (Y.R. Gel).

1 Supported in part by National Natural Science Foundation of China NSFC, 11631003, NSFC 11690012.
2 Supported in part by National Social Science Foundation of China, 19CTJ013.
3 Supported in part by the National Science Foundation, USA grants DMS 1736417, DMS 1925346, ECCS 1824710 and IIS 1633331.

https://doi.org/10.1016/j.jmva.2021.104732
0047-259X/

progress of data acquisition technology, the collection of such relational information becomes realistically feasible, thereby making it easier to integrate this critical aspect into the classification process.

In turn, relational data, formed by a collection of entities and connections among them can be described using the concept of complex networks [3,33]. Here, nodes represent, for example, observed individuals, with a set of the associated node attributes, for instance, socio-demographic information, and edges correspond to relationships between individuals. Among the most recent classification approaches for such relational data are machine learning tools, e.g., embedding approaches [12], neural networks [39] and network-based regularization algorithms [4,24], which have gained a particular popularity in bioinformatics and health care applications [34]. In addition, various logistic regression techniques, e.g., network-based logistic regression model, are widely adopted as one of the primary benchmark approaches in statistical sciences and machine learning for addressing the classification problem for relational data [40]. Intuitively, such relational data and the associated resulting complex networks are likely to be both heterogeneous and exhibit a highly nontrivial geometric structure, for instance, driven by various hidden communities and substructures that are not detectable with conventional Euclidian-based metrics. Nevertheless, despite an ever increasing interest in systematic assessment of network dependence in classification and prediction tasks and a growing evidence of the key role of data shape in organization of complex systems, most existing classification tools for relational data tend to neglect the intrinsic geometry of the observed relational data.

In this paper we introduce the concept of data depth to classification of relational data which we describe as a complex network with multiple node attributes. Data depth is a nonparametric data-driven method that systematically accounts for the underlying data geometry by assigning an ordering to each data point with respect to its position within a given data cloud or probability distribution [26,42]. A higher value of a data depth implies a higher centrality in the data cloud. The depth contour with such a natural center-outward ordering serves as a topological map of the data. As a result, clusters and outliers can be then evaluated simultaneously in a quick and visual manner. In addition to statistical sciences, data depth is rapidly gaining its popularity in machine learning and data sciences due to its wide applicability in classification, anomaly detection, and data visualization, for overview see, e.g., [15,17,19,25,37] and references therein. Despite a high utility of depth methods in multivariate and functional data analysis, data depth remains largely under-explored in complex network analysis. Among recent efforts in this direction are unsupervised depth-based clustering of graphs [7,35], depth-based classification (without node attributes) [36], and depth-based analysis of a random sample of graphs following a probability model on the space of all graphs of a given size [10].

Our goal in this paper is to explore utility of data depth to classification of relational data in a form of a complex network with multiple node attributes. We propose a probabilistic model for the observed data that consists of two parts. First, a logistic regression model is employed to describe relationships between the class label and attributes of each node. Second, we introduce a network model where the link probability between any two nodes is assumed to depend on their class labels and data depths of the nodes in the respective classes. The above two parts constitute the depth-based network classification model (DNC). The parameters of the DNC can then be estimated by a maximum likelihood method within a logistic regression framework [14]. Third, the approximate prediction rules of the DNC are obtained using approximate simplification from posterior probability of the unknown class label. Finally, we derive asymptotic properties of DNC as the network order tends to infinity under some mild conditions.

The rest of the paper is organized as follows. Section 2 introduces the definition of data depth, the DNC model and its corresponding approximate prediction rule, and the asymptotic classification theory. In Section 3, the proposed geometrically-enhanced classification approach is validated on a broad range of synthetic data and is illustrated in application to user classification of the one of the largest Chinese social media, Sina Weibo. The paper concludes in Section 4 with a discussion and future research directions. All theoretical derivations are presented in the Appendix.

2. Depth-based network classification

In this section we introduce the proposed data depth methodology to classification of relational data, starting from the overview of data depth concepts (Section 2.1), and followed by the modeling framework (Section 2.2), the approximate prediction rule (Section 2.3) of DNC and the asymptotic classification theory (Section 2.4) for the new depth-based network classification approach.

2.1. Background on data depth

In the last two decades, a concept of data depth is shown to be an attractive nonparametric tool to analyze multivariate data without making prior assumptions about underlying probability distributions. Recently, data depth has witnessed a new momentum in statistics, data science and machine learning due to its high utility in high dimensional, functional and categorical data analysis.

A data depth is a function that measures how closely an observed point $x \in \mathbb{R}^d$, $d \geq 2$, is located to the "center" of a finite set $\mathcal{X} \subset \mathbb{R}^d$, or relative to F, a probability distribution in $\mathbb{R}^d$. Data depth measures the "depth" (or "outlyingness") of a given object or a set of objects with respect to an observed data cloud. A higher value of a data depth implies a deeper location, or higher centrality in the data cloud.

Given their role in descriptive analysis, depth-based approaches also have broad applications in classification and clustering problems, as well as in visualization of the detected clusters, high-dimensional and functional data studies,

usually conjunct with microarray gene expression and other biomedical applications, see [5,18,23,31] and references therein. Depth-based approaches provide two main advantages, that is, such tools are more robust against outliers and are intrinsically geometric. The intuitive idea of depth function is to better capture a probabilistic geometry of the underlying data, that is, the position of each point with respect to the whole data cloud and figure out a possibility that a certain point belongs to the community, given a geometric structure of this community.

Depth functions vary in terms of computational complexity, robustness, and reflection of particular data cloud properties [16,26,30,32,42]. Choice of the most feasible depth function largely depends on particular desired properties in a given study, such as robustness, behavior of the depth outside of the convex support, and computational speed. Choice of a given depth function may also depend on the desired properties to be achieved. For example, the more robust depth such as projection depth could be used if the main goal is to enhance robustness against potential outliers. The geometric depth such as simplicial depth or half-space depth could be used if no additional information about the data is available, as geometric depths usually capture more accurately the true underlying geometric structure. In general, there exists no systematic rule which depth function is the most appropriate for a particular analytic task and a given data cloud. Choice of depth function may be dictated by some desirable properties to be achieved and often constitutes a trial-and-error approach based on a cross-validation.

Definition 1. Let $D(\cdot,\cdot)$ be a bounded, nonnegative mapping from $\mathbb{R}^d \times \mathcal{F}$ to $\mathbb{R}$. If $D(\cdot,\cdot)$ satisfies the following four properties, (i) Affine invariance: $D(Ax+b, F_{Ax+b}) = D(x, F_x)$ holds for any random vector X, constant vector b and any nonsingular matrix A; (ii) Maximality at center: $D(\theta, F) = sup_{x\in\mathbb{R}^d} D(x, F)$ holds for any distribution F having center θ; (iii) Monotonicity relative to deepest point: for any distribution F having deepest point θ, $D(x, F) \le D(\theta + \alpha(x-\theta), F)$ holds for $\alpha \in [0, 1]$; (iv) Vanishing at infinity: $D(x, F)$ tends to 0 as $\|x\|$ tends to infinity, then $D(\cdot, F)$ is called the statistical depth function [42].

We consider three nonnegative and bounded depth functions which are one of the most widely used depths across a broad range of analytic problems:

- (Mahalanobis depth (MhD))

  $$MhD_F(x) = \{1 + (x-\mu_F)'\Sigma_F^{-1}(x-\mu_F)\}^{-1},$$

  where μ_F and Σ_F are the mean vector and covariance matrix of F, respectively. The Mahalanobis depth [30,42] measures the outlyingness of the point with respect to the center of the distribution, and allows to handle the elliptical family of distributions easily, including a Gaussian case. However, the Mahalanobis depth is less robust and fails to distinguish two distributions which share first two moments.
- (Random Projection depth (RPD))

  $$RPD_F(x) = \{1 + \sup_{\|u\|=1} |u'x - Med(F_u)|/MAD(F_u)\}^{-1},$$

  where F_u is the distribution of $u'X$, Med(F_u) is the median of F_u, $MAD(F_u)$ is the median absolute deviation of F_u. Random projection depth stochastically approximates projection depth which also measures the outlyingness of the point with respect to the deepest point of the distribution. RPD is robust against possible extreme observations [41].
- (Tukey depth (TD))

  $$TD_F(x) = \inf_H \{P_F(H)\},$$

  where H is a closed half-space in $\mathbb{R}^d$ and $x \in H$, measures the tailedness of the point with respect to the deepest point of the distribution F. For the discussion on computationally efficient algorithms of for the Tukey depth in higher dimensions, see [6,8].

In general, RPD and TD tend to be appropriate for any distribution. When the distribution is symmetric, MhD might be a preferred choice due to its lower computational costs comparing to the other two depths.

2.2. The proposed model framework of DNC

Let $Y_i \in \{0, 1\}$ be the binary response (class label) of the ith $(1 \le i \le n)$ observation, and $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$ be the associated p-dimensional attributes, $Y = (Y_1, \ldots, Y_n)^\top \in \{0, 1\}^n$ and $X = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times p}$. The adjacency matrix of the network is then defined as $A_n = (a_{ij}) \in \{0, 1\}^{n\times n}$, where $a_{ij} = 1$ if node i follows node j, otherwise $a_{ij} = 0$. Furthermore, we assume that $a_{ii} = 0$ for any $1 \le i \le n$. Note that the subscript n of A_n indicates the order of the network and the asymmetry of A_n is common in practice. For example, the fact that one user in Sina Weibo follows another user does not mean the converse is true.

Now we turn to describing the relationship among X, Y and A_n. First, we model the relationship between X and Y, assume that Y_i-s are conditionally independent given X_i-s and are modeled via a logistic regression model (LR)

$$\ln\left\{\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}\right\} = X_i^\top \beta, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the regression coefficient. Assuming that a_{ij}-s are conditionally independent given X and Y, we then consider an LR model

$$\ln\left\{\frac{P(a_{ij}=1|X_i,X_j,Y_i,Y_j)}{1-P(a_{ij}=1|X_i,X_j,Y_i,Y_j)}\right\} = \omega_{Y_iY_j} + \phi^\top D^{ij},$$

where intercept $\omega_{Y_iY_j}$ only depends on the class labels of Y_i and Y_j. Denote F and G as the distribution of class 1 and class 0, then

$$D^{ij} = (D(X_i,F), D(X_i,G), D(X_j,F), D(X_j,G))^\top,$$

where $D(\cdot,\cdot)$ is a user-selected data depth function defined in Section 2.1. Since F and G are unknown, we consider the following approximate model

$$\ln\left\{\frac{P(a_{ij}=1|X,Y)}{1-P(a_{ij}=1|X,Y)}\right\} = \omega_{Y_iY_j} + \phi^\top D_n^{ij}. \tag{2}$$

Now, denote $\omega = (\omega_{11}, \omega_{10}, \omega_{01}, \omega_{00})^\top \in \mathbb{R}^4$ and

$$D_n^{ij} = (D(X_i,F_n), D(X_i,G_n), D(X_j,F_n), D(X_j,G_n))^\top,$$

where F_n is the empirical distribution of $\{X_i : Y_i = 1, 1 \le i \le n\}$, i.e., class 1, G_n is the empirical distribution of $\{X_i : Y_i = 0, 1 \le i \le n\}$, i.e., class 0. As a result, model (1) and model (2) constitute the DNC model. Note that, vector D_n^{ij} indicates depths of node i and j in class 1 and 0, which incorporates both the covariates information and the class information. $\phi = (\phi_1, \phi_2, \phi_3, \phi_4)^\top \in \mathbb{R}^4$ are the regression parameters. In particular, if $\phi = 0$, then model (2) degenerates to the block model, which implies that the link formation only depends on the class labels of nodes. However, the block model ignores the impact of covariates on link formation. Hence, the term $\phi^\top D_n^{ij}$ is appended to integrate the covariates information in model (2). Moreover, the statistical significance of ω and ϕ in real data analysis in Section 3.3 illustrates the reasonability of the model settings in application.

Let $\theta = (\beta^\top, \phi^\top, \omega^\top)^\top$ be a vector of model parameters. Then, the likelihood function can be represented as

$$\begin{aligned}\mathcal{L}(\theta) &= P(Y, A_n|X) = \prod_{i=1}^n P(Y_i|X_i) \prod_{i\neq j} P(a_{ij}|Y,X) \\ &= \prod_{i=1}^n \left\{\frac{\exp(X_i^\top\beta)}{1+\exp(X_i^\top\beta)}\right\}^{Y_i} \left\{\frac{1}{1+\exp(X_i^\top\beta)}\right\}^{1-Y_i} \\ &\quad \times \prod_{i\neq j} \left\{\frac{\exp(\omega_{Y_iY_j} + \phi^\top D_n^{ij})}{1+\exp(\omega_{Y_iY_j} + \phi^\top D_n^{ij})}\right\}^{a_{ij}} \left\{\frac{1}{1+\exp(\omega_{Y_iY_j} + \phi^\top D_n^{ij})}\right\}^{1-a_{ij}}.\end{aligned}$$

The maximum likelihood estimator (MLE) is then $\hat\theta = (\hat\beta^\top, \hat\phi^\top, \hat\omega^\top)^\top = \arg\max_\theta \mathcal{L}(\theta)$, which can be obtained from two separate logistic regression models (1) and (2).

2.3. Prediction rule of DNC

Without loss of generality, denote a new node as $n+1$, our goal is to predict the unknown class label Y_{n+1} based on Y, X, X_{n+1} and A_{n+1}, i.e., the adjacency matrix of $n+1$ nodes, which depends on the posterior probability $P(Y_{n+1}|Y, X, X_{n+1}, A_{n+1})$. Since D_{n+1} cannot be acquired when Y_{n+1} is unknown, we substitute D_{n+1} with D_n and derive the approximate prediction rule. To illustrate the performance of the approximation and the theoretical results in the consequent sections, we need the following technical assumptions.

Assumption C1 (*Parameter Boundedness*)**.** There exist some finite positive constants M_x and M_β, such that $\|X_i\|_1 \le M_X$ for any $1 \le i \le n$, and $\|\beta\|_1 \le M_\beta$.

Assumption C2 (*Network Sparsity*)**.** There exist positive constants c_{kl} and $\gamma < 1$, such that $\omega_{kl} = c_{kl} - \gamma \ln n$, where $k, l \in \{0, 1\}$.

Assumption C1 is the condition about the range of X and parameter β, which is trivial under a fixed covariates dimension. As for Assumption C2, the assumption is equivalent to $\pi_{ij}^{Y_iY_j} = O_p(n^{-\gamma})$, where $\pi_{ij}^{Y_iY_j} = P(a_{ij}=1|Y,X)$. Note that Assumption C2 is intrinsically related to the concept of network sparsity. That is, a network is called sparse if the network density, i.e., $\{n(n-1)\}^{-1}\sum_{i\neq j}^n a_{ij}$, tends to 0 as the network order n tends to infinity [20]. Hence, by Assumption C2, $\mathrm{E}(a_{ij}) = O(n^{-\gamma})$, which implies that the network density tends to zero as n tends to infinity.

Now, we can derive the posterior distribution $P(Y_{n+1}|Y, X, X_{n+1}, A_{n+1})$ and obtain the following approximate prediction rule

$$\ln\left\{\frac{P(Y_{n+1}=1|Y,X,X_{n+1},A_{n+1})}{P(Y_{n+1}=0|Y,X,X_{n+1},A_{n+1})}\right\} \approx X_{n+1}^{\top}\beta + Q_1 - Q_2 + Q_3 - Q_4, \tag{3}$$

where terms Q_ts $(1 \le t \le 4)$ are defined as,

$$Q_1 = \sum_{i=1}^{n}\Big[a_{i(n+1)}(\omega_{Y_i 1} + \phi^{\top} D_n^{i(n+1)}) - \ln\big\{1 + \exp(\omega_{Y_i 1} + \phi^{\top} D_n^{i(n+1)})\big\}\Big],$$

$$Q_2 = \sum_{i=1}^{n}\Big[a_{i(n+1)}(\omega_{Y_i 0} + \phi^{\top} D_n^{i(n+1)}) - \ln\big\{1 + \exp(\omega_{Y_i 0} + \phi^{\top} D_n^{i(n+1)})\big\}\Big],$$

$$Q_3 = \sum_{j=1}^{n}\Big[a_{(n+1)j}(\omega_{1 Y_j} + \phi^{\top} D_n^{(n+1)j}) - \ln\big\{1 + \exp(\omega_{1 Y_j} + \phi^{\top} D_n^{(n+1)j})\big\}\Big],$$

$$Q_4 = \sum_{j=1}^{n}\Big[a_{(n+1)j}(\omega_{0 Y_j} + \phi^{\top} D_n^{(n+1)j}) - \ln\big\{1 + \exp(\omega_{0 Y_j} + \phi^{\top} D_n^{(n+1)j})\big\}\Big].$$

Here, data depths $D_n^{i(n+1)} = (D(X_i, F_n), D(X_i, G_n), D(X_{n+1}, F_n), D(X_{n+1}, G_n))^{\top}$ and $D_n^{(n+1)j} = (D(X_{n+1}, F_n), D(X_{n+1}, G_n), D(X_j, F_n), D(X_j, G_n))^{\top}$ are calculated based on F_n and G_n. The detailed derivation of (3) and discussion on feasibility of approximation are provided in the Appendix. In applications, unknown parameters in (3) are replaced by the respective MLE $\hat{\theta}$. Note that the resulting approximate prediction rule depends on both the network structure and the data depth of nodes in the two classes.

2.4. Theoretical properties

We now turn to deriving theoretical properties of DNC.

Theorem 1. *For DNC, under Assumptions C1 and C2, if $0 < \gamma < 1/3$, then $P\big(\hat{Y}_{n+1} = 1|Y_{n+1} = 1, X, X_{n+1}\big) \to 1$ as $n \to \infty$, where $\hat{Y}_{n+1}$ is the predicted class label of node $n+1$ according to the approximate prediction rule of DNC.*

Theorem 1 implies that when the network provides enough information, i.e., $0 < \gamma < 1/3$, the prediction accuracy of DNC tends to 1 as $n \to \infty$. Note that, the bigger the value of γ is, the sparser the network is. In addition, by the result of Theorem 1, simple algebra shows that the misclassification probability $P(\hat{Y}_{n+1} \neq Y_{n+1}|X)$ in DNC tends to 0 as the network order $n \to \infty$ as long as the probabilities of any class do not degenerate to zero. The proof of Theorem 1 is given in the Appendix.

3. Numerical studies

3.1. Simulation models and performance measurements

We investigate robustness and classification performance of the newly proposed DNC approach on synthetic data in a broad range of finite sample scenarios, that is, our simulation settings include continuous, discrete and mixed (contains both continuous and discrete variables) distributions, unbalanced designs and a case of outliers. In particular, we consider six simulation models, where $X_i = (X_{i1}, \ldots, X_{ip})^{\top}$ $(1 \le i \le n)$ is distributed according to one of the models below and the predictor dimension is fixed to be $p = 5$:

- Model 1: X_i is simulated from a multivariate normal distribution with mean $(0, 0, 0, 0, 0)^{\top}$ and covariance $\Sigma_X = (\sigma_{j_1 j_2}) \in \mathbb{R}^{5\times 5}$ $(1 \le j_1, j_2 \le 5)$, where $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$ [38].
- Model 2: X_i is simulated from a multivariate normal distribution with mean $(1, 3, 1, 0.5, 1)^{\top}$ and the covariance as in Model 1. Model 2 considers the case of unbalanced classes.
- Model 3: X_i is simulated from two multivariate normal distributions with probability 0.98 and 0.02, and corresponds to a case with outliers. The means are $(0, 0, 0, 0, 0)^{\top}$ and $(4, 4, 4, 4, 4)^{\top}$ for the first and second multivariate normal distributions, accordingly. Covariances of multivariate normal distributions are the same as in Model 1.
- Model 4: X_{ij} $(1 \le j \le 5)$ is simulated from uniform distribution $U(0, 1)$.
- Model 5: X_{ij} $(1 \le j \le 5)$ is simulated from binomial distribution $B(1, p_j)$, where $p_1 = 0.5$, $p_2 = 0.6$, $p_3 = 0.7$, $p_4 = 0.6$ and $p_5 = 0.5$.
- Model 6: X_{i1} and X_{i2} are simulated from $B(1, 0.5)$ and $B(1, 0.6)$ separately, and $(X_{i3}, X_{i4}, X_{i5})^{\top}$ is simulated from multivariate normal distribution with mean $(0, 0, 0)^{\top}$ and covariance $\Sigma_X = (\sigma_{j_1 j_2}) \in \mathbb{R}^{3\times 3}$, where $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$ for $1 \le j_1, j_2 \le 3$.

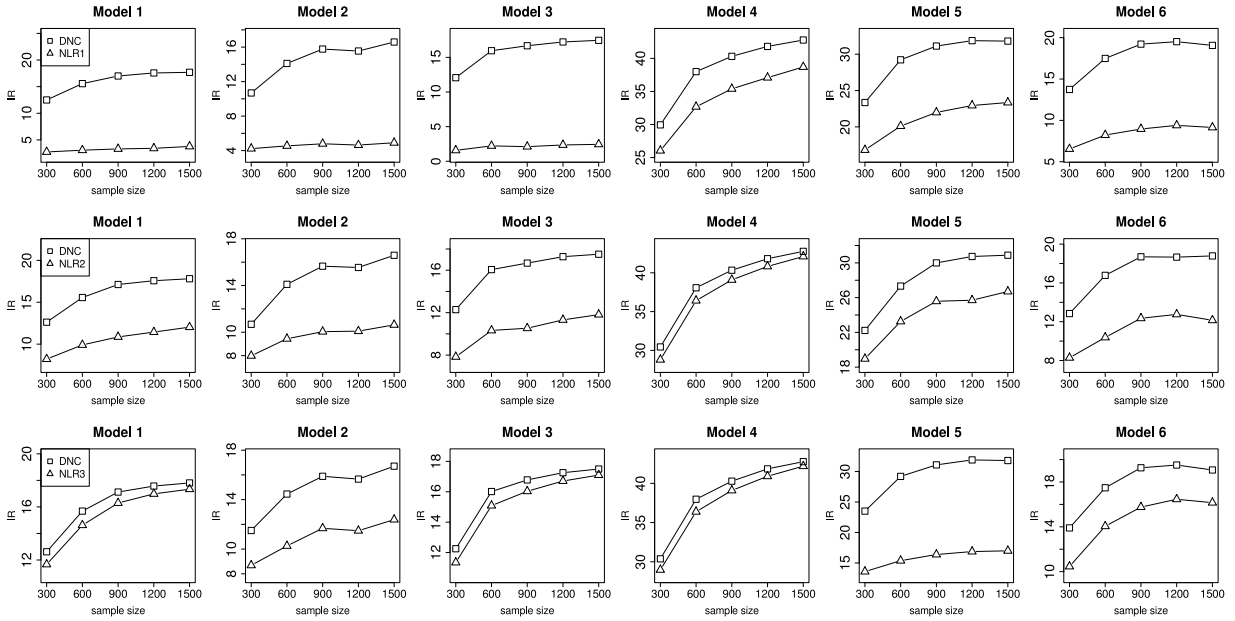

Fig. 1. The classification performance of DNC and NLR in respect to LR for the six simulation models, measured in terms of IR values, where $IR = (AUC_{DNC}/AUC_{LR} - 1) * 100\%$. The 1st row is the result of NLR and DNC with MhD, the 2nd row is the result of NLR and DNC with RPD, the 3rd row is the result of NLR and DNC with TD.

Next, the class label Y_i is generated according to model (1), and the adjacency matrix A_n is generated according to model (2). We set $\beta = (-1, 0.8, 1, -2, 1)^\top$, $\omega = (-0.35, -0.5, -0.5, -0.35)^\top$ and $\phi = (2.5, -2.5, 1.5, -1.5)^\top$.

For each simulation model, we set five sample sizes, i.e., $n = 300, 600, 900, 1200, 1500$. The number of replications S is 100. We consider three depth functions, namely, the Mahalanobis depth (MhD), the random projection depth (RPD) and the Tukey depth (TD).

We measure accuracy of parameter estimation in terms of the root mean squared error $\text{RMSE}_\beta = (S^{-1}\sum_{s=1}^{S}\|\hat{\beta}^{(s)} - \beta\|^2)^{1/2}$, $\text{RMSE}_\phi = (S^{-1}\sum_{s=1}^{S}\|\hat{\phi}^{(s)} - \phi\|^2)^{1/2}$ and $\text{RMSE}_\omega = (S^{-1}\sum_{s=1}^{S}\|\hat{\omega}^{(s)} - \omega\|^2)^{1/2}$. Note that the coefficient β in the LR model is as same as β in (1) of the proposed DNC models, since estimation of (1) and (2) is separable.

Finally, to evaluate the prediction accuracy in out-of-sample settings, we generate another n_0 of 500 observations, indexed by $n + 1, \ldots, n + 500$ in each replication as follows. In the sth replication, after generating the predictor $X_{n+i}^{(s)}$ $(1 \le i \le 500)$ for each simulation model, the class label $Y_{n+i}^{(s)}$ $(1 \le i \le 500)$ is derived according to (1), and the network links between this particular testing sample and the existing n training subjects are generated by (2). Note that an index $r_1 = S^{-1}\sum_{s=1}^{S} r_1^{(s)}$ is provided to illustrate the case of unbalanced classes, where $r_1^{(s)}$ is the proportion of class 1 in training set for the sth replication. We compare performance of DNC in respect to LR which ignores the impact of the relational structure throughout classification process and to a network-based logistic regression model (NLR) [40] which similarly to the DNC approach explicitly accounts for relational effects in the observed data. Comparison of DNC to NLR allows to assess the contribution of underlying geometric properties of the data into classification performance.

The area under the receiver operating characteristic curve (AUC) value is used to derive index $IR = (AUC_{DNC}/AUC_{LR} - 1) * 100\%$, which evaluates the prediction accuracy improvement ratio of DNC and NLR with respect to LR. The detailed simulation results are summarized in Figs. 1–2 and Table 1.

3.2. Simulation results

As Fig. 1 indicates, the IR values are positive for all sample sizes and depth functions for each model, implying that the AUC values for the DNC and NLR models are greater than the respective LR values in all considered scenarios. As sample sizes increase, the IR values tend to increase, implying a higher gain in performance delivered by DNC and NLR in respect to LR for higher observational sample size. The results show that incorporating network structure indeed improves the classification accuracy. Moreover the DNC outperforms NLR in all cases, which indicates the advantage of using data depth. Performance gains of DNC appear to be highest, i.e., highest IR values, in Models 4 and 5, corresponding to uniform and binomial distributions.

As we can see from Fig. 2, among the three considered depth functions, MhD and TD tend to deliver stable performance across all six models, and the yielded prediction accuracy among MhD and TD is comparable. In turn, RPD appears to be less competitive than MhD and TD, especially for Model 5, i.e., binomial distribution, and Model 6, i.e., the mixed distribution,

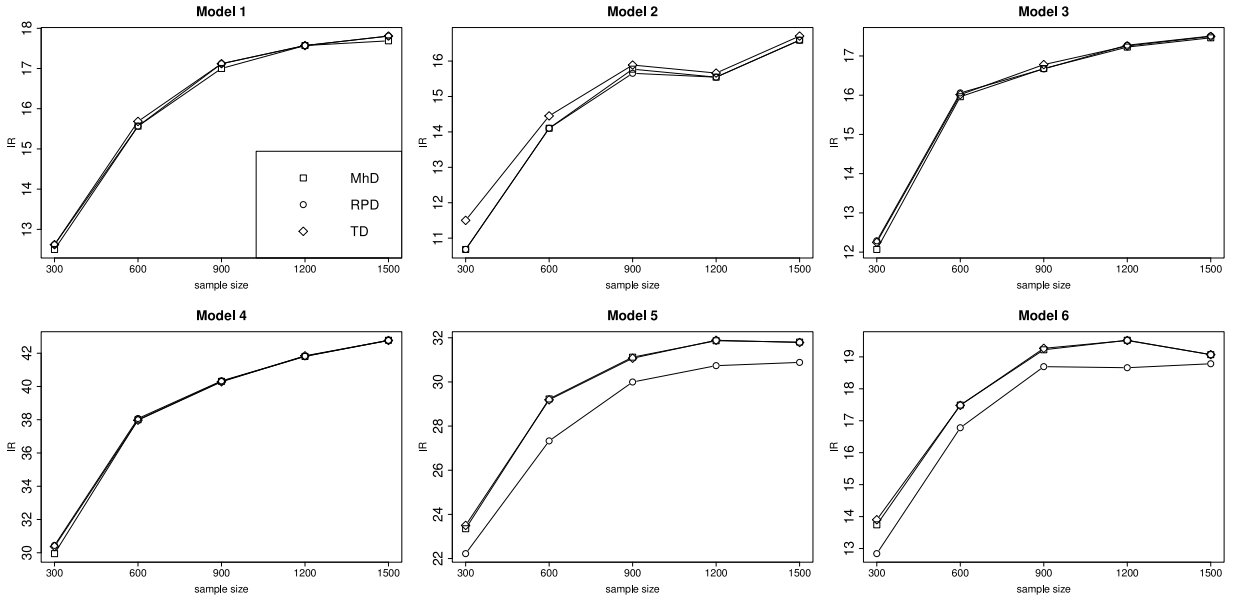

Fig. 2. The classification performance of DNC with three depth functions, i.e., MhD, RPD and TD, in respect to LR for the six simulation models, measured in terms of IR values, where $IR = (AUC_{DNC}/AUC_{LR} - 1) * 100\%$.

Table 1
The RMSE of parameters estimation results as well as the unbalanced index r_1 for six simulation models with three depth functions, i.e., MhD, RPD and TD, for sample size from 300 to 1500 are given.

	n	r_1	RMSE						
			β	ϕ_{MhD}	ϕ_{RPD}	ϕ_{TD}	ω_{MhD}	ω_{RPD}	ω_{TD}
Model 1	300	0.577	0.496	0.329	0.528	0.661	0.074	0.119	0.101
	600	0.503	0.330	0.156	0.239	0.325	0.041	0.064	0.052
	900	0.507	0.258	0.114	0.185	0.247	0.025	0.042	0.038
	1200	0.512	0.256	0.093	0.133	0.153	0.021	0.032	0.028
	1500	0.501	0.222	0.078	0.126	0.117	0.019	0.028	0.024
Model 2	300	0.870	0.574	0.373	0.588	0.801	0.114	0.175	0.125
	600	0.835	0.366	0.193	0.266	0.386	0.048	0.089	0.065
	900	0.839	0.324	0.134	0.199	0.251	0.037	0.055	0.046
	1200	0.833	0.269	0.103	0.172	0.170	0.028	0.039	0.036
	1500	0.819	0.232	0.092	0.158	0.134	0.023	0.036	0.030
Model 3	300	0.537	0.482	0.309	0.435	0.718	0.077	0.107	0.091
	600	0.507	0.349	0.165	0.255	0.335	0.034	0.050	0.050
	900	0.518	0.290	0.116	0.185	0.222	0.028	0.037	0.044
	1200	0.514	0.243	0.093	0.144	0.155	0.021	0.030	0.030
	1500	0.510	0.203	0.080	0.126	0.116	0.018	0.027	0.026
Model 4	300	0.537	0.898	0.630	0.926	0.628	0.064	0.126	0.101
	600	0.492	0.626	0.352	0.495	0.333	0.038	0.075	0.056
	900	0.478	0.492	0.220	0.355	0.234	0.026	0.048	0.044
	1200	0.485	0.467	0.180	0.275	0.174	0.017	0.033	0.027
	1500	0.487	0.397	0.155	0.267	0.156	0.014	0.028	0.026
Model 5	300	0.497	0.580	0.659	2.337	0.531	0.260	0.629	0.732
	600	0.528	0.451	0.373	2.403	0.308	0.150	0.696	0.447
	900	0.480	0.341	0.290	2.348	0.226	0.139	0.675	0.319
	1200	0.498	0.300	0.221	2.331	0.180	0.109	0.643	0.247
	1500	0.523	0.258	0.172	2.234	0.146	0.082	0.628	0.223
Model 6	300	0.523	0.527	0.490	1.793	0.682	0.104	0.576	0.200
	600	0.512	0.337	0.215	1.471	0.345	0.057	0.495	0.108
	900	0.528	0.285	0.173	1.415	0.274	0.044	0.539	0.088
	1200	0.483	0.248	0.124	1.762	0.211	0.032	0.540	0.066
	1500	0.505	0.204	0.101	1.338	0.157	0.026	0.548	0.050

Table 2
The estimation of the model (2) for the Sina Weibo data set, where $\theta = (\omega_{00}, \omega_{11} - \omega_{00}, \omega_{10} - \omega_{00}, \omega_{01} - \omega_{00}, \phi_1, \phi_2, \phi_3, \phi_4)^\top$, " $***$ " indicates p-value ≤ 0.001 and " $**$ " indicates $0.001 < p$-value ≤ 0.01.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
Estimate	−3.67	1.62	0.16	0.03	2.26	−2.39	−0.07	0.51
S.E.	0.01	0.02	0.01	0.01	0.05	0.07	0.05	0.06
p-value	$***$	$***$	$**$	$***$	$***$	$***$	0.176	$***$

which both contain discrete covariates. From Model 1 to Model 4, all three data depth functions deliver similar accuracy gains over the benchmark LR.

We now turn to assessing estimation accuracy for parameters β, ϕ and ω based on the data depth approach. As Table 1 implies, the RMSE values for β, ϕ and ω decrease as n increases under all simulation settings except for the RPD case, which is corresponding to the results in Fig. 2. The most accurate estimations of ϕ and ω tend to be delivered by the MhD. In particular, for Model 1 RMSE values for ϕ decrease from 0.329 to 0.078, i.e., improvement of 76% for n of 300 and 1500, respectively; and RMSE values for ω decrease from 0.074 for n of 300 to 0.019 for n of 1500, i.e., improvement of 74%. The results are similar across other five models, except that the most accurate estimator of ϕ is delivered by the TD in Model 5, which suggests the TD may be a preferred estimator in the discrete predictor case. Finally, the obtained finite sample performance for Model 2 is comparable to other models implying robustness of the depth-based method for a case of unbalanced classes ($r_1 > 0.8$ in Model 2).

These findings indicate that incorporating relational information and accounting for its underlying geometric structure can lead to substantial gains in classification performance and robust conclusions in a broad range of scenarios.

3.3. Classification of the Sina Weibo users

We now illustrate application of the proposed depth-based approach to classification of users of Sina Weibo which is the one of the largest online social media platforms in China. Sina Weibo allows different users to follow each other and share various types of information between the connected users. Our data set contains 2000 users. Each user has multiple attributes which are authentication (1 for authenticated users and 0 otherwise), number of Weibo posts, number of personal labels, number of days the user has been registered, number of personal collections, gender (1 for male and 0 for female), and the Master of Business Administration (MBA) community label (1 for being a member of the MBA community and 0 otherwise). Here, we regard the MBA attribute as class label, and our goal is to predict whether a Weibo user has an MBA degree based on his/her connections, i.e., the followers and followees of each user, and personal attributes. Accurate user classification assists the social platform in developing more efficient recommender systems and to implement personalized marketing strategy. In this case study, incorporating the relational information, i.e., a network structure is intuitive, as the MBA users of Weibo are more likely to share similar online behavior and personal preferences; in turn, adopting a data depth may assist in recovering latent geometric patterns among Weibo MBA users, for instance, hidden community substructures of users who are alumni of the same University, employed within the same industry sector, or share other professional interests — that is, geometrically enhanced depth-based classification can assist in revealing information which is not explicitly provided in the observed data set.

To test utility of the proposed DNC models with MhD, RPD and TD functions, we randomly split the data set into two parts. The first 1000 observations are regarded as training set and the remaining 1000 observations are used for testing. We compare the AUC values for the DNC with competing LR and NLR models on the testing set over 200 random splits. In addition, we compare the AUC values with the results delivered by support vector machine method (SVM) and random forest method (RF).

Fig. 3 illustrates the obtained results. In particular, we find that the DNC approach with all the considered depth functions, i.e., MhD, RPD, and TD, outperforms its competitors. First, the median AUC values delivered by methods with network information are greater than 0.76, while the AUC values of methods without using network are less than 0.72. Second, the average AUC values of the network contained methods are 0.746 (NLR), 0.754 (DNC with MhD), 0.749 (DNC with RPD) and 0.756 (DNC with TD), which shows that the DNC method tends to outperform NLR.

In summary, we can conclude that incorporation of dependency among entities and geometrically enhanced data depth analysis into the classification process may lead to substantial gains in prediction accuracy in relational data. The model fitting results for (2) are summarized in Table 2, and show that most model coefficients are found to be statistically significant, i.e., the p-value of all coefficients except θ_7 is less than 0.05, which also reflect contributions of the underlying relational and geometric information to the classification process.

4. Discussion

We have proposed a new probabilistic model, i.e., depth-based network classification model (DNC), which integrates dependency among observations and accounts for intrinsic data geometric structure via a data depth concept, into classification tasks for relational data. More specifically, we have modeled the observed relational data as a complex

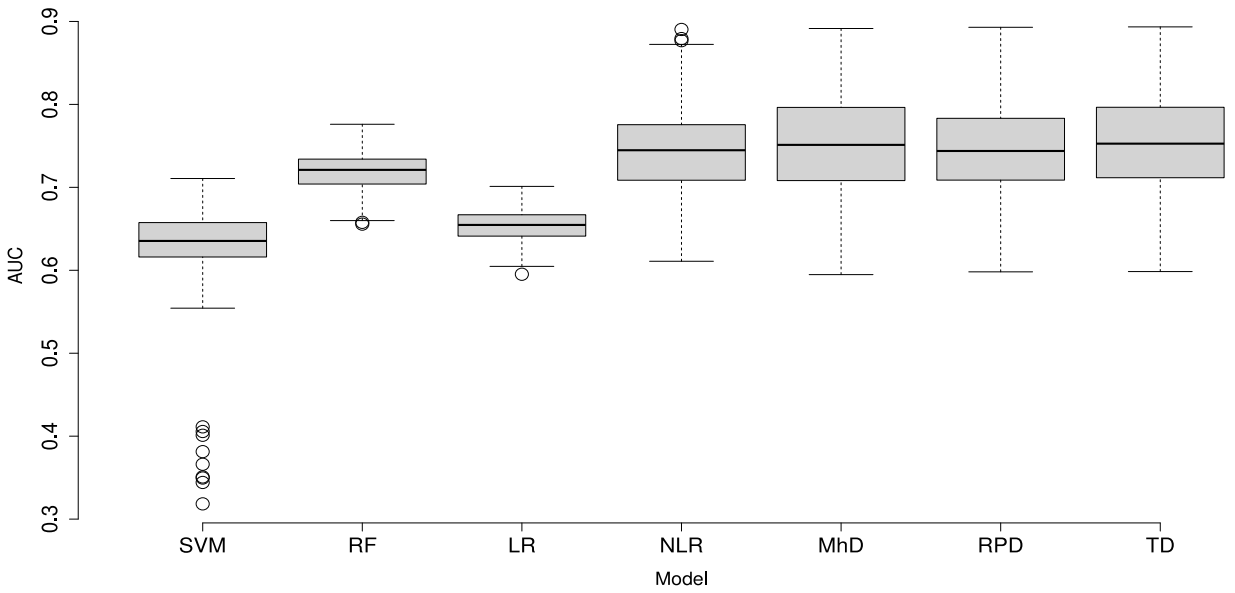

Fig. 3. Prediction performance of DNC with three depth functions, i.e., MhD, RPD and TD, and four competitors (SVM, RF, LR and NLR) for the Sina Weibo user classification.

network with multiple node attributes. The relationship between the class label and attributes of each node is described by logistic regression (LR) model, and the link probability between any two nodes is assumed to be dependent on their class labels and the data depths of the nodes in two classes. The proposed new geometrically-enhanced classification method has shown to outperform the benchmarks approaches such as LR which ignores the impact of the relational structure throughout classification process and the network-based logistic regression model (NLR) [40] which similarly to the DNC approach explicitly accounts for relational effects in the observed data, in all considered finite sample scenarios which include continuous, discrete and mixed type distributions, unbalanced designs and a case of outliers. Furthermore, the DNC method has delivered competitive performance in predicting membership labels in a binary classification task of the Sina Weibo users — that is, the one of the largest Chinese online social media platforms.

In the future, we plan to advance the proposed data depth approach to multi-class classification beyond LR and integrate DNC with tree-based methods and neural networks. Furthermore, the DNC models can be extended to an unsupervised case, i.e., community detection based on geometric structure and data shape of node attributes [9,21]. Finally, we will evaluate utility of the proposed DNC approaches for risk scoring in blockchain transaction graphs, particularly, in conjunction with whale detection and anti-money laundering efforts [1,2].

CRediT authorship contribution statement

Xu Zhang: Methodology, Formal analysis, Visualization, Software, Funding acquisition. **Yahui Tian:** Methodology, Formal analysis, Validation. **Guoyu Guan:** Conceptualization, Methodology, Investigation, Funding acquisition. **Yulia R. Gel:** Conceptualization, Methodology, Investigation, Funding acquisition.

Appendix

Detailed derivation of (3). First, we introduce the continuity of depth functions, i.e., MhD, RPD and TD, on the distribution function, without loss of generality, denoted as F, which is useful in the derivations to follow. The continuity of a depth function on F means that, for any fixed x, $D(x, F) - D(x, F_n) = o_p(1)$, if F_n converges to F, where F_n is empirical distribution function. For MhD and TD, by [30], we have for any fixed x, $MhD_F(x) - MhD_{F_n}(x) = o_p(1)$ and $TD_F(x) - TD_{F_n}(x) = o_p(1)$, as F_n converges to F. As for RPD, by [41], we have $RPD_F(x) - RPD_{F_n}(x) = o_p(1)$, as F_n converges to F, under some mild conditions.

Second, the posterior probability of Y_{n+1} given all observed information is as follows

$$\begin{aligned} P(Y_{n+1}|Y, X, X_{n+1}, A_{n+1}) &= \{P(Y, A_{n+1}|X, X_{n+1})\}^{-1} P(Y, Y_{n+1}, A_{n+1}|X, X_{n+1}) \\ &\propto P(Y, Y_{n+1}|X, X_{n+1}) P(A_{n+1}|Y, Y_{n+1}, X, X_{n+1}) \end{aligned}$$

$$= \prod_{i=1}^{n+1} P(Y_i|X_i) \prod_{i\neq j}^{n+1} P(a_{ij}|Y, Y_{n+1}, X, X_{n+1})$$
$$\propto \frac{\{\exp(X_{n+1}^\top \beta)\}^{Y_{n+1}}}{1+\exp(X_{n+1}^\top \beta)} \prod_{i\neq j}^{n+1} \frac{\{\exp(\omega_{Y_iY_j} + \phi^\top D_{n+1}^{ij})\}^{a_{ij}}}{1+\exp(\omega_{Y_iY_j} + \phi^\top D_{n+1}^{ij})}.$$

Here symbol "$\propto$" means "be proportional to". That is, the part which is independent on Y_{n+1} is omitted. Hence,

$$\ln\{P(Y_{n+1}|Y, X, X_{n+1}, A_{n+1})\} \propto Y_{n+1}X_{n+1}^\top \beta - \ln\{1+\exp(X_{n+1}^\top \beta)\}$$
$$+ \sum_{i\neq j}^{n+1}\Big[a_{ij}(\omega_{Y_iY_j} + \phi^\top D_{n+1}^{ij}) - \ln\{1+\exp(\omega_{Y_iY_j} + \phi^\top D_{n+1}^{ij})\}\Big] \doteq H(Y_{n+1}, D_{n+1}).$$

Note that because Y_{n+1} is unknown, we cannot acquire the quantity D_{n+1}^{ij}. To fill the gap, we substitute D_{n+1}^{ij} with D_n^{ij}, which results in the approximation as follows

$$H(Y_{n+1}, D_{n+1}) \approx Y_{n+1}X_{n+1}^\top \beta - \ln\{1+\exp(X_{n+1}^\top \beta)\}$$
$$+ \sum_{i\neq j}^{n+1}\Big[a_{ij}(\omega_{Y_iY_j} + \phi^\top D_n^{ij}) - \ln\{1+\exp(\omega_{Y_iY_j} + \phi^\top D_n^{ij})\}\Big] \doteq H(Y_{n+1}, D_n).$$

Hence, under the assumption of continuity of data depth, the impact of approximation on the prediction rule is negligible.

Finally, we obtain the following approximate prediction rule

$$\ln\left\{\frac{P(Y_{n+1}=1|Y, X, X_{n+1}, A_{n+1})}{P(Y_{n+1}=0|Y, X, X_{n+1}, A_{n+1})}\right\}$$
$$\approx X_{n+1}^\top \beta + \sum_{i=1}^{n}\Big[a_{i(n+1)}(\omega_{Y_i1} + \phi^\top D_n^{i(n+1)}) - \ln\{1+\exp(\omega_{Y_i1} + \phi^\top D_n^{i(n+1)})\}\Big]$$
$$- \sum_{i=1}^{n}\Big[a_{i(n+1)}(\omega_{Y_i0} + \phi^\top D_n^{i(n+1)}) - \ln\{1+\exp(\omega_{Y_i0} + \phi^\top D_n^{i(n+1)})\}\Big]$$
$$+ \sum_{j=1}^{n}\Big[a_{(n+1)j}(\omega_{1Y_j} + \phi^\top D_n^{(n+1)j}) - \ln\{1+\exp(\omega_{1Y_j} + \phi^\top D_n^{(n+1)j})\}\Big]$$
$$- \sum_{j=1}^{n}\Big[a_{(n+1)j}(\omega_{0Y_j} + \phi^\top D_n^{(n+1)j}) - \ln\{1+\exp(\omega_{0Y_j} + \phi^\top D_n^{(n+1)j})\}\Big].$$

Proof of Theorem 1. We divide this proof into two steps as follows.

Step 1. According to the approximate prediction rule (3), the $n+1$th node will be labeled as 1 if $X_{n+1}^\top\beta + Q_1 - Q_2 + Q_3 - Q_4 > 0$, which is equivalent to

$$g(\theta) = X_{n+1}^\top \beta + \sum_{i=1}^{n} a_{i(n+1)} \ln \frac{\pi_{i(n+1)}^{Y_i 1}}{\pi_{i(n+1)}^{Y_i 0}} + \sum_{i=1}^{n}\{1-a_{i(n+1)}\}\ln\frac{1-\pi_{i(n+1)}^{Y_i 1}}{1-\pi_{i(n+1)}^{Y_i 0}}$$
$$+ \sum_{j=1}^{n} a_{(n+1)j} \ln \frac{\pi_{(n+1)j}^{1Y_j}}{\pi_{(n+1)j}^{0Y_j}} + \sum_{j=1}^{n}\{1-a_{(n+1)j}\}\ln\frac{1-\pi_{(n+1)j}^{1Y_j}}{1-\pi_{(n+1)j}^{0Y_j}} > 0,$$

where $\theta = (\beta^\top, \phi^\top, \omega^\top)^\top$. Furthermore, $g(\theta)$ can be rewritten as

$$X_{n+1}^\top \beta + \sum_{i=1}^{n}\left[a_{i(n+1)}Y_i \ln\frac{\pi_{i(n+1)}^{11}}{\pi_{i(n+1)}^{10}} + \{1-a_{i(n+1)}\}Y_i \ln\frac{1-\pi_{i(n+1)}^{11}}{1-\pi_{i(n+1)}^{10}}\right]$$
$$+ \sum_{i=1}^{n}\left[a_{i(n+1)}(1-Y_i)\ln\frac{\pi_{i(n+1)}^{01}}{\pi_{i(n+1)}^{00}} + \{1-a_{i(n+1)}\}(1-Y_i)\ln\frac{1-\pi_{i(n+1)}^{01}}{1-\pi_{i(n+1)}^{00}}\right]$$

$$+\sum_{i=1}^{n}\left[a_{(n+1)i}Y_i\ln\frac{\pi^{11}_{(n+1)i}}{\pi^{01}_{(n+1)i}}+\{1-a_{(n+1)i}\}Y_i\ln\frac{1-\pi^{11}_{(n+1)i}}{1-\pi^{01}_{(n+1)i}}\right]$$
$$+\sum_{i=1}^{n}\left[a_{(n+1)i}(1-Y_i)\ln\frac{\pi^{10}_{(n+1)i}}{\pi^{00}_{(n+1)i}}+\{1-a_{(n+1)i}\}(1-Y_i)\ln\frac{1-\pi^{10}_{(n+1)i}}{1-\pi^{00}_{(n+1)i}}\right].$$

Hence, the probability $P(\hat{Y}_{n+1}=1|Y_{n+1}=1,X,X_{n+1})$ is equal to $P\big(g(\hat{\theta})>0|Y_{n+1}=1,X,X_{n+1}\big)$. Denote $g(\hat{\theta})=X^{\top}_{n+1}\hat{\beta}+F_1+F_2+F_3+F_4$, where

$$F_1=\sum_{i=1}^{n}\left[a_{i(n+1)}Y_i\ln\frac{\hat{\pi}^{11}_{i(n+1)}}{\hat{\pi}^{10}_{i(n+1)}}+\{1-a_{i(n+1)}\}Y_i\ln\frac{1-\hat{\pi}^{11}_{i(n+1)}}{1-\hat{\pi}^{10}_{i(n+1)}}\right],$$
$$F_2=\sum_{i=1}^{n}\left[a_{i(n+1)}(1-Y_i)\ln\frac{\hat{\pi}^{01}_{i(n+1)}}{\hat{\pi}^{00}_{i(n+1)}}+\{1-a_{i(n+1)}\}(1-Y_i)\ln\frac{1-\hat{\pi}^{01}_{i(n+1)}}{1-\hat{\pi}^{00}_{i(n+1)}}\right],$$
$$F_3=\sum_{i=1}^{n}\left[a_{(n+1)i}Y_i\ln\frac{\hat{\pi}^{11}_{(n+1)i}}{\hat{\pi}^{01}_{(n+1)i}}+\{1-a_{(n+1)i}\}Y_i\ln\frac{1-\hat{\pi}^{11}_{(n+1)i}}{1-\hat{\pi}^{01}_{(n+1)i}}\right],$$
$$F_4=\sum_{i=1}^{n}\left[a_{(n+1)i}(1-Y_i)\ln\frac{\hat{\pi}^{10}_{(n+1)i}}{\hat{\pi}^{00}_{(n+1)i}}+\{1-a_{(n+1)i}\}(1-Y_i)\ln\frac{1-\hat{\pi}^{10}_{(n+1)i}}{1-\hat{\pi}^{00}_{(n+1)i}}\right].$$

Thus, the proof is equivalent to prove that $g(\hat{\theta})$ tends to infinity in probability one, as $n\to\infty$. By [29] and Assumption C1, we know that $X^{\top}_{n+1}\hat{\beta}=X^{\top}_{n+1}(\hat{\beta}-\beta)+X^{\top}_{n+1}\beta$ is bounded in probability. Because the mathematical forms of F_1, F_2, F_3 and F_4 are similar, we only need to prove F_1 tends to positive infinity in probability one as $n\to\infty$.

Step 2. First, denote $p_i=\exp(x_i^{\top}\beta)/\{1+\exp(x_i^{\top}\beta)\}$, hence the boundedness of X and β guarantees the boundedness of p_i. Write $F_1=F_1-F_0+F_0$, where

$$F_0=\sum_{i=1}^{n}p_i\left[\pi^{11}_{i(n+1)}\ln\frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}}+\{1-\pi^{11}_{i(n+1)}\}\ln\frac{1-\pi^{11}_{i(n+1)}}{1-\pi^{10}_{i(n+1)}}\right].$$

Next, by Assumption C2, it is easy to derive that there exist some finite positive constants $\nu<1/3$, γ, such that $\pi^{kl}_{ij}\propto n^{-\gamma}$, $\nu n^{-\gamma}\le\pi^{kl}_{ij}\le 1-\nu n^{-\gamma}$ for $k,l\in\{0,1\}$. Then with Assumption C1 and the result of lemma 3 in Guan et al. 2018 [11], we have

$$F_0\ge\frac{3}{2}\sum_{i=1}^{n}p_i\nu^2n^{-2\gamma}\ge\frac{3}{2}min\{p_i\}\nu^2n^{1-2\gamma}=O(n^{1-2\gamma}).$$

Next, we need to derive the order of $|F_1-F_0|$. First, we have

$$F_1-F_0=\sum_{i=1}^{n}\left\{a_{i(n+1)}Y_i\ln\frac{\hat{\pi}^{11}_{i(n+1)}}{\hat{\pi}^{10}_{i(n+1)}}-p_i\pi^{11}_{i(n+1)}\ln\frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}}\right\}$$
$$+\sum_{i=1}^{n}\left[\{1-a_{i(n+1)}\}Y_i\ln\frac{1-\hat{\pi}^{11}_{i(n+1)}}{1-\hat{\pi}^{10}_{i(n+1)}}-p_i\{1-\pi^{11}_{i(n+1)}\}\ln\frac{1-\pi^{11}_{i(n+1)}}{1-\pi^{10}_{i(n+1)}}\right]\doteq H_1+H_2,$$

where

$$H_1=\sum_{i=1}^{n}\left\{a_{i(n+1)}Y_i\ln\frac{\hat{\pi}^{11}_{i(n+1)}}{\hat{\pi}^{10}_{i(n+1)}}-p_i\pi^{11}_{i(n+1)}\ln\frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}}\right\},$$
$$H_2=\sum_{i=1}^{n}\left[\{1-a_{i(n+1)}\}Y_i\ln\frac{1-\hat{\pi}^{11}_{i(n+1)}}{1-\hat{\pi}^{10}_{i(n+1)}}-p_i\{1-\pi^{11}_{i(n+1)}\}\ln\frac{1-\pi^{11}_{i(n+1)}}{1-\pi^{10}_{i(n+1)}}\right].$$

Next,

$$H_1=\sum_{i=1}^{n}\left\{a_{i(n+1)}Y_i\ln\frac{\hat{\pi}^{11}_{i(n+1)}}{\hat{\pi}^{10}_{i(n+1)}}-a_{i(n+1)}Y_i\ln\frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}}\right\},$$
$$+\sum_{i=1}^{n}\left\{a_{i(n+1)}Y_i\ln\frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}}-p_i\pi^{11}_{i(n+1)}\ln\frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}}\right\}$$

$$= \sum_{i=1}^{n} a_{i(n+1)} Y_i \left\{ \ln \frac{\hat{\pi}^{11}_{i(n+1)}}{\hat{\pi}^{10}_{i(n+1)}} - \ln \frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}} \right\} + \sum_{i=1}^{n} \{a_{i(n+1)} Y_i - p_i \pi^{11}_{i(n+1)}\} \ln \frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}} \doteq H_{11} + H_{12}.$$

For H_{11}, by Assumptions C1 and C2, the asymptotic normality properties of logistic regression and the Δ-method, we have

$$\hat{\pi}^{11}_{i(n+1)} - \pi^{11}_{i(n+1)} \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma = \{\pi^{11}_{i(n+1)}(1 - \pi^{11}_{i(n+1)})\}^2 z_i^\top \{\sum_{i=1}^n \pi^{11}_{i(n+1)}(1 - \pi^{11}_{i(n+1)}) z_i z_i^\top\}^{-1} z_i$, and $z_i = (\omega_{11}, \phi_1, \phi_2, \phi_3, \phi_4)^\top$. Hence $|\hat{\pi}^{11}_{i(n+1)} - \pi^{11}_{i(n+1)}| = O_p(n^{-\frac{\gamma+1}{2}})$. Moreover,

$$\left| \ln \hat{\pi}^{11}_{i(n+1)} - \ln \pi^{11}_{i(n+1)} \right| = \left| \ln\{1 + \frac{\hat{\pi}^{11}_{i(n+1)} - \pi^{11}_{i(n+1)}}{\pi^{11}_{i(n+1)}}\} \right| \le \sum_{k=1}^{\infty} \frac{1}{k} \left| \frac{\hat{\pi}^{11}_{i(n+1)} - \pi^{11}_{i(n+1)}}{\pi^{11}_{i(n+1)}} \right|^k = O_p\left(\sum_{k=1}^{\infty} \frac{1}{k} n^{k(\gamma-1)} \right) = O_p(n^{\frac{\gamma-1}{2}}).$$

Hence,

$$\left| \ln \frac{\hat{\pi}^{11}_{i(n+1)}}{\hat{\pi}^{10}_{i(n+1)}} - \ln \frac{\pi^{11}_{i(n+1)}}{\pi^{10}_{i(n+1)}} \right| = \left| \ln \hat{\pi}^{11}_{i(n+1)} - \ln \pi^{11}_{i(n+1)} - \ln \hat{\pi}^{10}_{i(n+1)} + \ln \pi^{10}_{i(n+1)} \right| \le \left| \ln \hat{\pi}^{11}_{i(n+1)} - \ln \pi^{11}_{i(n+1)} \right| + \left| \ln \hat{\pi}^{10}_{i(n+1)} - \ln \pi^{10}_{i(n+1)} \right| = O_p(n^{\frac{\gamma-1}{2}}).$$

For $\sum_{i=1}^n a_{i(n+1)} Y_i$, given $Y_{n+1} = 1$, its conditional expectation is $\sum_{i=1}^n \pi^{11}_{i(n+1)} p_i = O_p(n^{1-\gamma})$, and its variance is $\sum_{i=1}^n [\pi^{11}_{i(n+1)} p_i - \{\pi^{11}_{i(n+1)} p_i\}^2] = O_p(n^{1-\gamma})$. As a result, we have $H_{11} = O_p(n^{\frac{1-\gamma}{2}})$.

For H_{12}, by Assumption C2, $\ln\{\pi^{11}_{i(n+1)} / \pi^{10}_{i(n+1)}\} = O_p(1)$. That is, $H_{12} = O_p(n^{\frac{1-\gamma}{2}})$. Hence, $H_1 = O_p(n^{\frac{1-\gamma}{2}})$. Similarly, we can derive that $H_2 = O_p(n^{\frac{1-\gamma}{2}})$. Now, we have $|F_1 - F_0| = O_p(n^{\frac{1-\gamma}{2}})$. Hence, compared to the order of F_0, if $0 < \gamma < 1/3$, we have $g(\hat{\theta})$ tends to ∞ with probability one, as $n \to \infty$, which completes the proof.

References

[1] C.G. Akcora, Y. Li, Y.R. Gel, M. Kantarcioglu, BitcoinHeist: Topological data analysis for ransomware prediction on the bitcoin blockchain, in: Proceedings of the 29th International Joint Conference on Artificial Intelligence, 2020, pp. 4439–4445.
[2] C.G. Akcora, S. Purusotham, Y.R. Gel, M. Krawiec-Thayer, M. Kantarcioglu, How to not get caught when you launder money on blockchain? 2020, arXiv:2010.15082.
[3] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: structure and dynamics, Phys. Rep. 424 (2006) 175–308.
[4] A. Cloninger, Prediction models for graph-linked data with localized regression, in: Wavelets and Sparsity XVII, Vol. 10394, 2017, pp. 176–185.
[5] J.A. Cuesta-Albertos, M. Febrero-Bande, M.O. de la Fuente, The DD^G-classifier in the functional setting, Test 26 (2017) 119–142.
[6] J. Cuesta-Albertos, A. Nieto-Reyes, The random Tukey depth, Comput. Statist. Data Anal. 52 (2008) 4979–4988.
[7] A. Dey, Y. Gel, H. Poor, Intentional islanding of power grids with data depth, in: Proceedings of the IEEE Workshop on Computational Advances in Multi-Sensor Adaptive Processing, CAMSAP2017, 2017, pp. 1–5.
[8] R. Dyckerhoff, P. Mozharovskyi, Exact computation of the halfspace depth, Comput. Statist. Data Anal. 98 (2016) 19–30.
[9] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[10] D. Fraiman, N. Fraiman, R. Fraiman, Nonparametric statistics of dynamic networks with distinguishable nodes, Test 26 (2017) 546–573.
[11] G. Guan, N. Shan, J. Guo, Feature screening for ultrahigh dimensional binary data, Stat. Interface 11 (2018) 41–50.
[12] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, Adv. Neural Inf. Process. Syst. (2017) 1024–1034.
[13] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science & Business Media, New York, 2009.
[14] D.W. Hosmer Jr, S. Lemeshow, R.X. Sturdivant, Applied Logistic Regression, John Wiley & Sons, New York, 2013.
[15] X. Huang, Y. Gel, CRAD: Clustering with robust autocuts and depth, in: Proceedings of the IEEE International Conference on Data Mining, ICDM, 2017, pp. 925–930.
[16] M. Hubert, P.J. Rousseeuw, S. Van Aelst, High-breakdown robust multivariate methods, Statist. Sci. 23 (2008) 92–119.
[17] M.-H. Jeong, Y. Cai, C.J. Sullivan, S. Wang, Data depth based clustering analysis, in: Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2016, pp. 1–10.
[18] R. Jörnsten, Clustering and classification based on the L_1 data depth, J. Multivariate Anal. 90 (2004) 67–89.
[19] M. Kleindessner, U. von Luxburg, Lens depth function and k-relative neighborhood graph: versatile tools for ordinal data analysis, J. Mach. Learn. Res. 18 (2017) 1889–1940.
[20] E.D. Kolaczyk, G. Csárdi, Statistical Analysis of Network Data with R, Vol. 65, Springer, New York, 2014.
[21] A. Lancichinetti, S. Fortunato, Community detection algorithms: a comparative analysis, Phys. Rev. E 80 (2009) 056117.
[22] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[23] J. Li, J.A. Cuesta-Albertos, R.Y. Liu, DD-classifier: nonparametric classification procedure based on DD-plot, J. Amer. Statist. Assoc. 107 (2012) 737–753.
[24] T. Li, E. Levina, J. Zhu, Prediction models for network-linked data, Ann. Appl. Stat. 13 (2019) 132–164.

[25] Y. Lil, U. AIslambekov, C. Akcora, E. Smirnova, Y.R. Gel, M. Kantarcioglu, Dissecting ethereum blockchain analytics: What we learn from topology and geometry of the ethereum graph? in: Proceedings of the 2020 SIAM International Conference on Data Mining, SDM, 2020, pp. 523–531.
[26] R.Y. Liu, J.M. Parelius, K. Singh, Multivariate analysis by data depth: descriptive statistics, graphics and inference, Ann. Statist. 27 (1999) 783–858.
[27] L. Luo, C. Chen, Z. Zhang, W.-J. Li, T. Zhang, Robust frequent directions with application in online learning, J. Mach. Learn. Res. 20 (2019) 1–41.
[28] X. Mai, R. Couillet, A random matrix analysis and improvement of semi-supervised learning for large dimensional data, J. Mach. Learn. Res. 19 (2018) 3074–3100.
[29] P. McCullagh, J. Nelder, Generalized Linear Models, Springer Science & Business Media, New York, 1989.
[30] K. Mosler, R. Hoberg, Data analysis and classification with the zonoid depth, DIMACS Ser. Discrete Math. Theoret. Comput. Sci. 72 (2006) 49–59.
[31] K. Mosler, P. Mozharovskyi, Fast DD-classification of functional data, Statist. Papers 58 (2017) 1055–1089.
[32] A. Nieto-Reyes, H. Battey, A topologically valid definition of depth for functional data, Statist. Sci. 31 (2016) 61–79.
[33] S.H. Strogatz, Exploring complex networks, Nature 410 (2001) 268–276.
[34] C. Su, J. Tong, Y. Zhu, P. Cui, F. Wang, Network embedding in biomedical data science, Brief. Bioinform. 21 (2020) 182–197.
[35] Y. Tian, Y.R. Gel, Fast community detection in complex networks with a k-depths classifier, Big and Complex Data Analysis (2017) 139–157.
[36] Y. Tian, Y.R. Gel, Fusing data depth with complex networks: Community detection with prior information, Comput. Statist. Data Anal. 139 (2019) 99–116.
[37] R.T. Whitaker, M. Mirzargar, R.M. Kirby, Contour boxplots: a method for characterizing uncertainty in feature sets from simulation ensembles, IEEE Trans. Vis. Comput. Graphics 19 (2013) 2713–2722.
[38] D.M. Witten, R. Tibshirani, Penalized classification using Fisher's linear discriminant, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (2011) 753–772.
[39] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. (2020) 4–24.
[40] X. Zhang, R. Pan, G. Guan, X. Zhu, H. Wang, Logistic regression with network structure, Statist. Sinica 30 (2020) 673–693.
[41] Y. Zuo, Projection-based depth functions and associated medians, Ann. Statist. 31 (2003) 1460–1490.
[42] Y. Zuo, R. Serfling, General notions of statistical depth function, Ann. Statist. 28 (2000) 461–482.