

کاربست یک الگوریتم جاسازی گراف برای جاسازی شبکه‌ی اندرکنش ناهمگن دارو-پروتئینی
برای پیش‌بینی دارو-هدف

فاطمه فتاحی

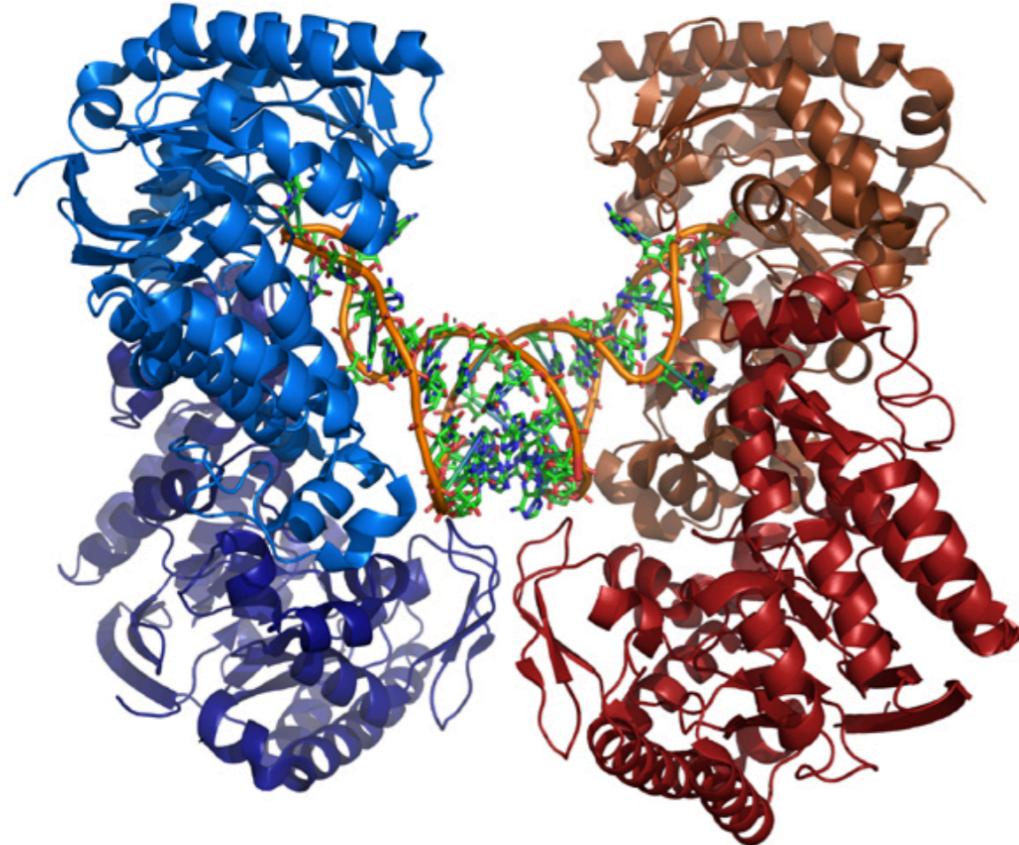
استاد راهنما: دکتر مینایی

1398 مهرماه

فهرست مطالب

- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

معرفی مسئله: مقدمه



• پیش بینی ارتباطات دارو-هدف

یک مرحله‌ی مهم در فرآیند کشف دارو و استفاده‌ی مجدد از دارو

• اهداف

تخصیص داروهای موثر جدید

کشف هدف جدیدی برای داروهای موجود

پیش بینی زودهنگام اثرات جانبی داروها

اثرگذاری داروهای موجود بر روی بیماری‌های دیگر که با نام استفاده‌ی مجدد از دارو شناخته می‌شود.

معرفی مسئله: مقدمه

• انواع روش‌ها برای حل مسئله

★ روش‌های آزمایشگاهی



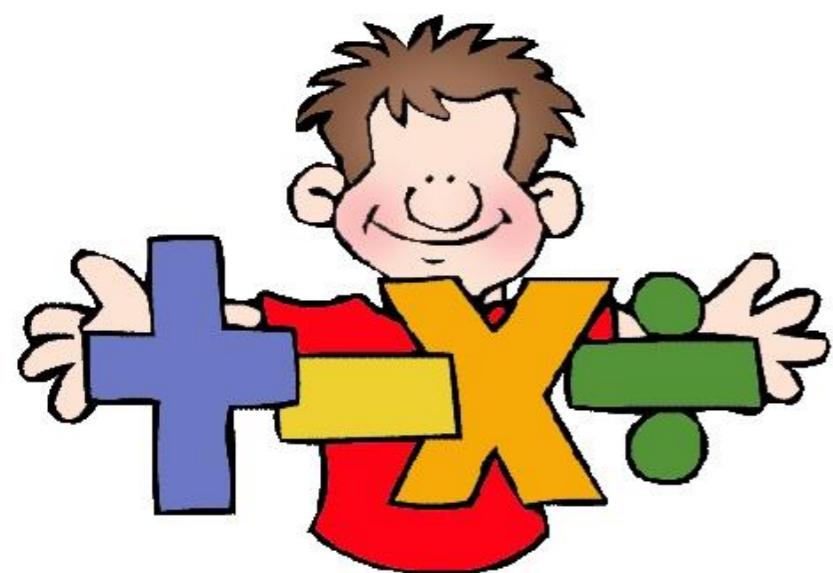
• زمان برو و بسیار پرهزینه (حدود 10-15 سال زمان و حداقل 800 میلیون دلار)

★ روش‌های محاسباتی

• کارآمد و موثر

★ انواع روش‌های محاسباتی

• روش‌های پیش‌بینی مبتنی بر شبکه



فهرست مطالب

- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

معرفی مسئله: شبکه (network)

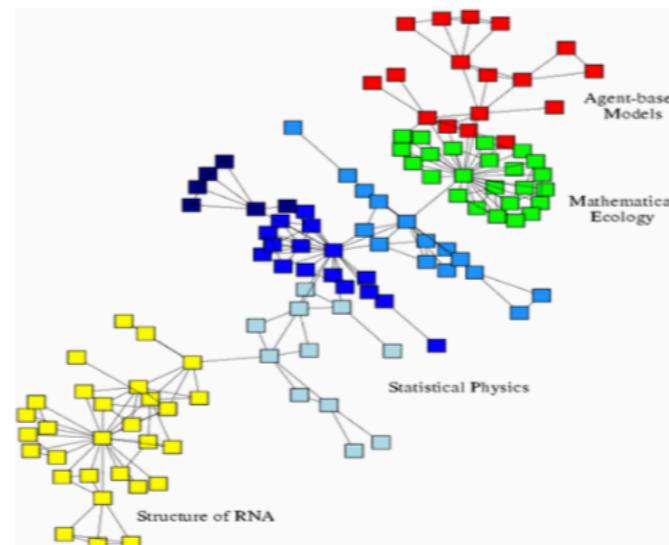
- شبکه ها یک زبان عمومی برای توصیف و مدل سازی سیستم های پیچیده هستند.

چرا از شبکه استفاده می کنیم؟ ★

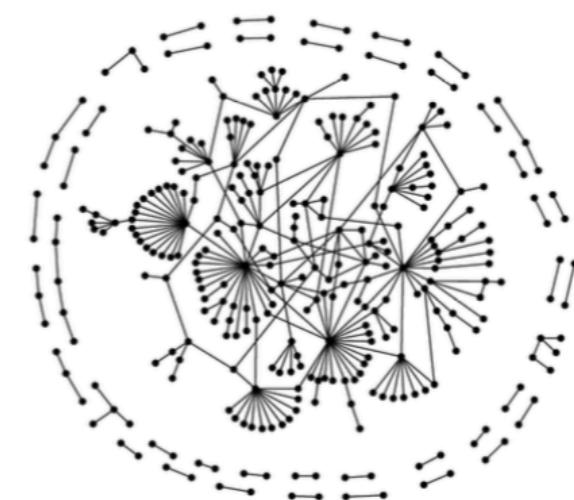
- به دلیل پیچیدگی و تنوع زیاد روابط و موجودیت ها در علم زیست پزشکی



Social networks



Economic networks

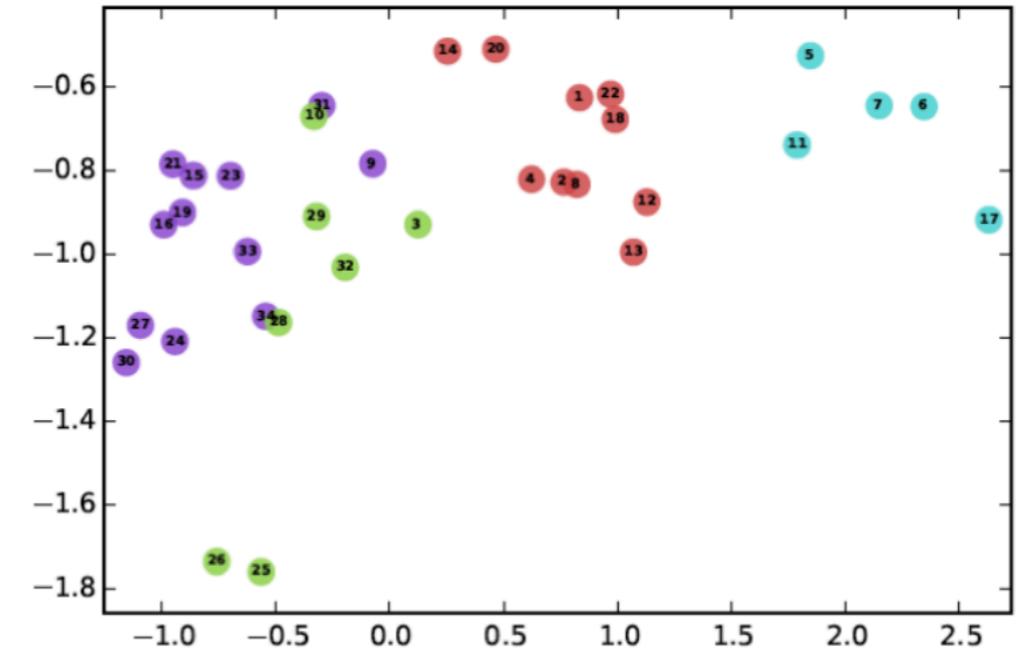
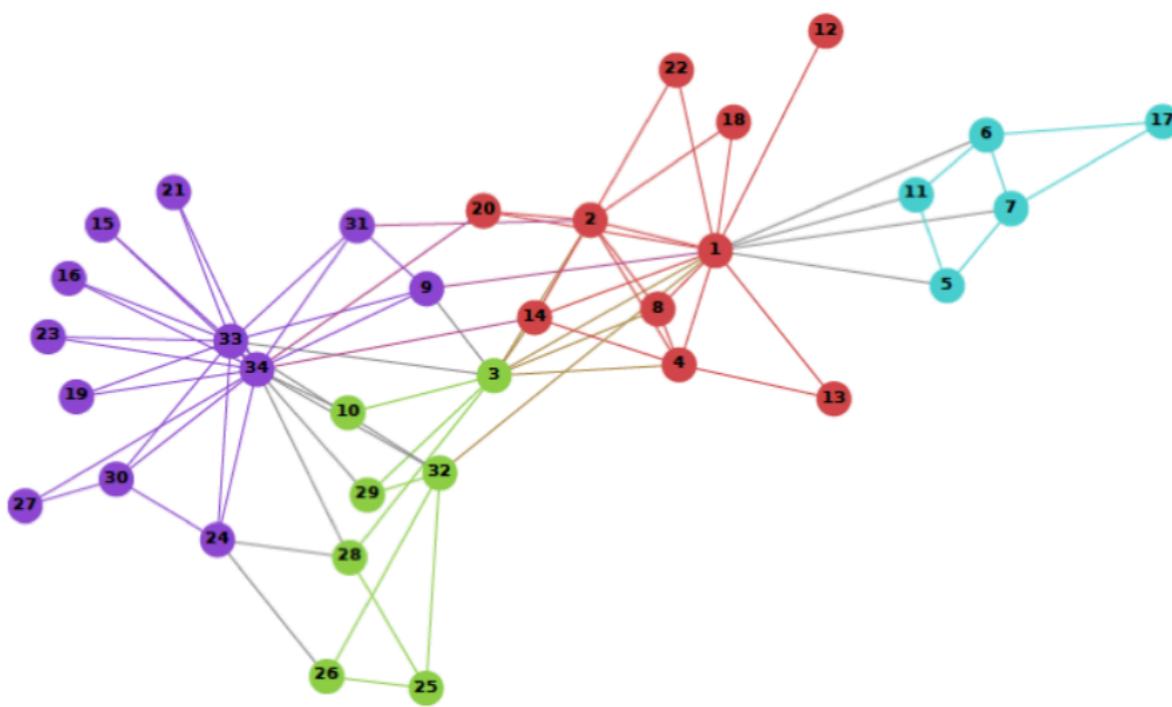


Biomedical networks

فهرست مطالب

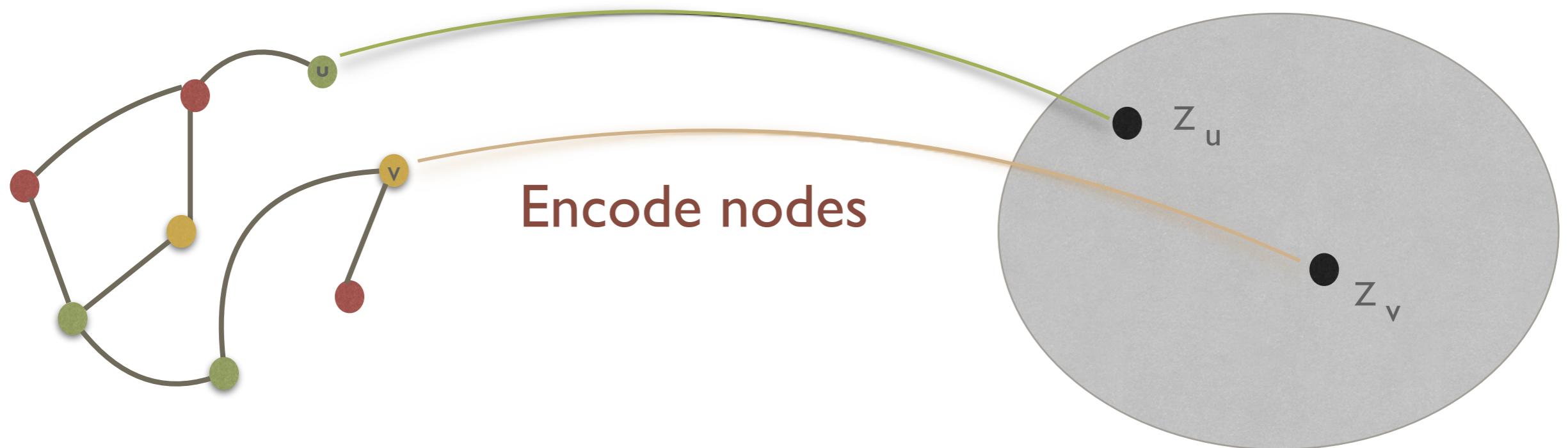
- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

معرفی مسئله: جاسازی نودها در گراف (node embedding)



جاسازی نودها به معنی تعبیه نودها در فضایی با ابعاد پایین‌تر است؛ به گونه‌ای که اطلاعات مربوط به آن‌ها از دست نزود. به این ترتیب گره‌های مشابه در گراف، در فضای جاسازی نیز به هم نزدیک هستند.

معرفی مسئله: جاسازی نودها در گراف (node embedding)



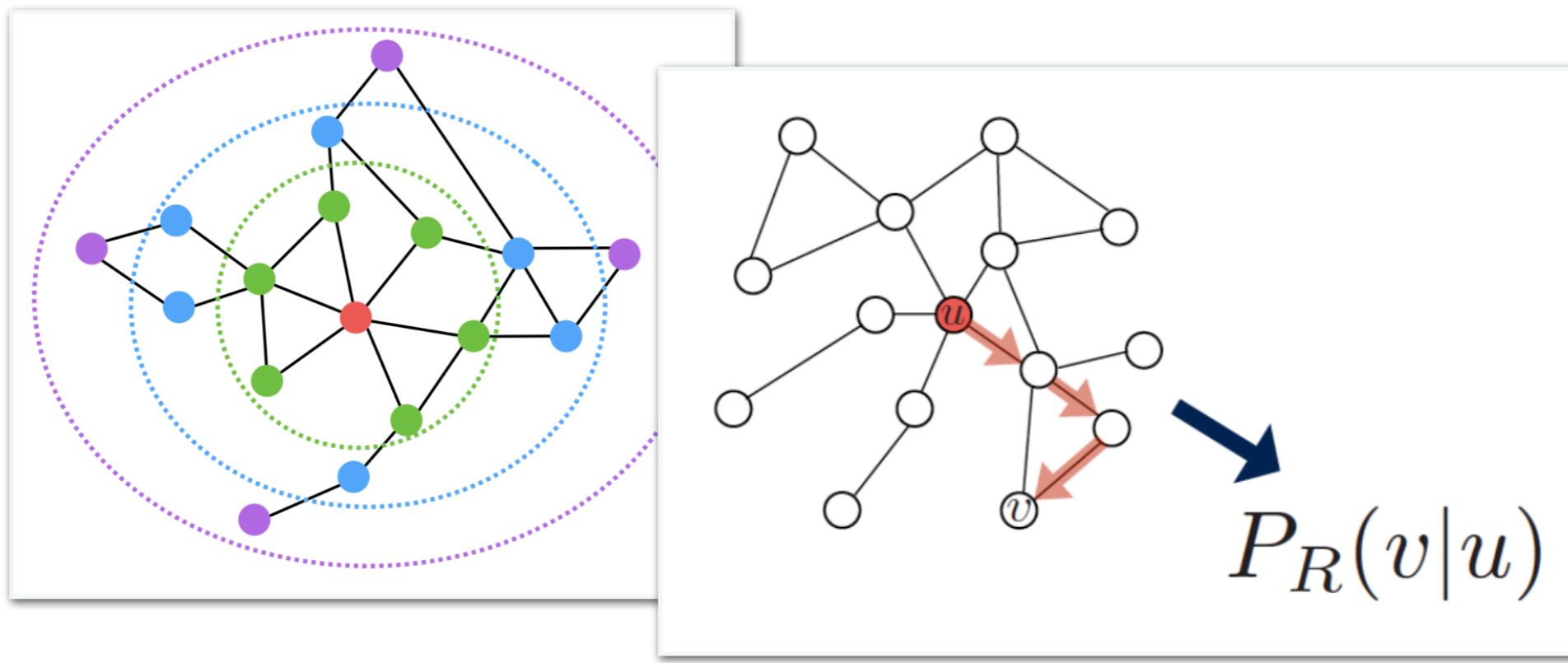
- هدف رمزگذاری گره‌ها به گونه‌ای است که شباهت در فضای تعبیه شده، شباهت در شبکه اصلی را نشان می‌دهد.

فهرست مطالب

- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

معرفی مسئله: انواع روش های تعیین شباهت نود (Node Similarity)

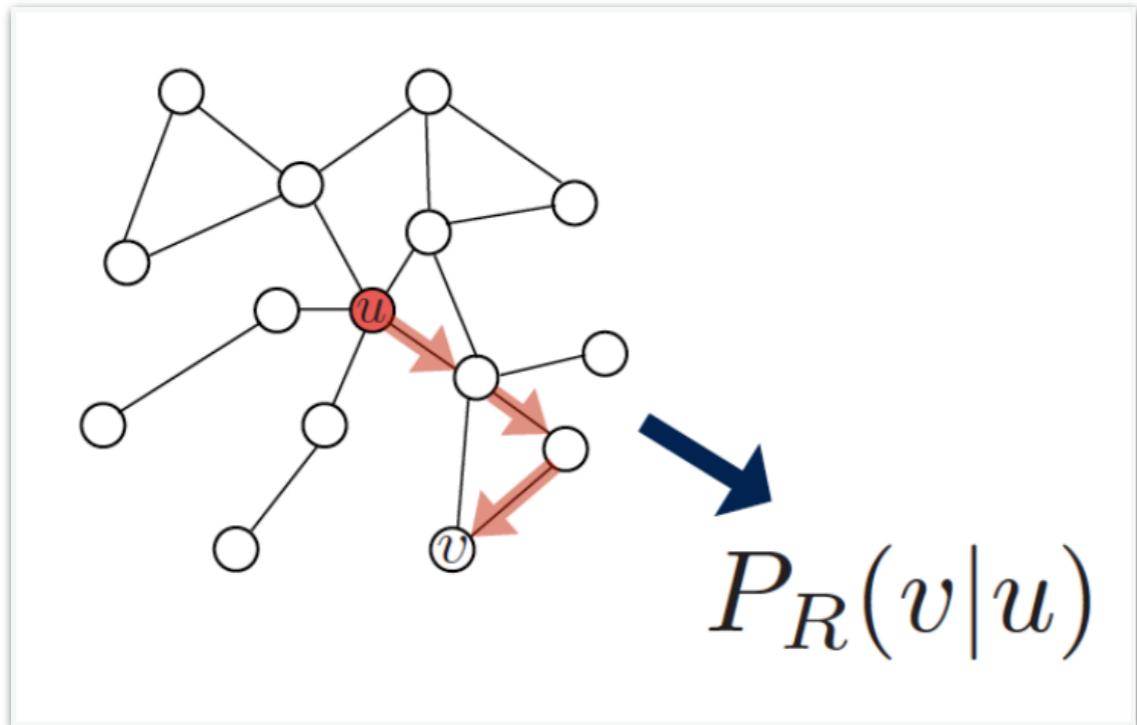
1. شباهت مبتنی بر وابستگی (Adjacency-based Similarity)
2. شباهت چند گامی (Multi-hop Similarity)
3. روش های پیاده روی تصادفی (Random Walk Approaches) ✓



معرفی مسئله: انواع روش های تعیین شباهت نود (Node Similarity)

★ روش پیاده روی تصادفی (Random Walk Approaches)

- احتمال ملاقات گره v در یک پیاده روی تصادفی که از گره u با استفاده از استراتژی R آغاز می شود:



- بهینه سازی جاسازی ها به منظور رمزگذاری پیاده روی های تصادفی

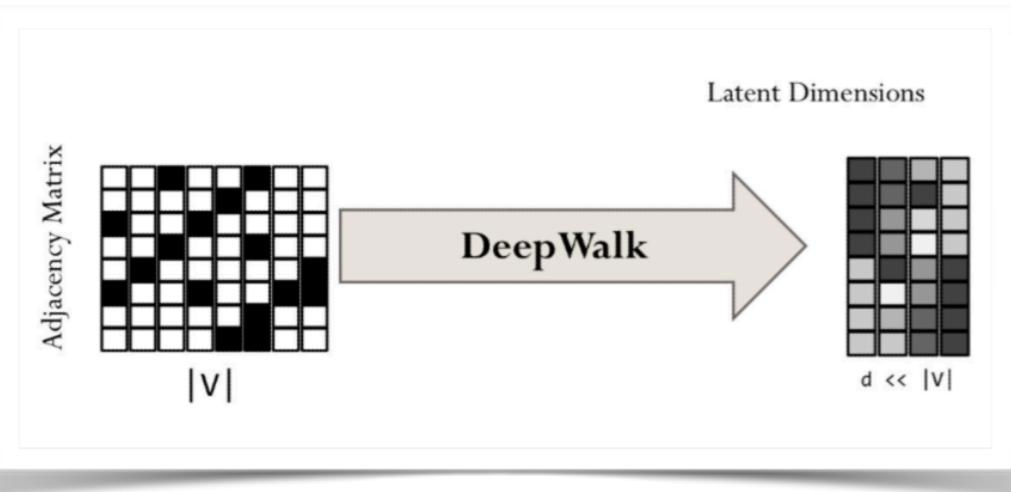
معرفی مسئله: انواع روش های تعیین شباخت نود (Node Similarity)

★ مزیت های روش های پیاده روی تصادفی

- **انعطاف پذیری:** علاوه بر اطلاعات همسایه های محلی شامل اطلاعات همسایه های مرتبه های بالاتر نیز می شود.
- **کارآیی:** به دلیل در نظر نگرفتن همه نودها در فرایند آموزش، پیچیدگی کمتری دارد.

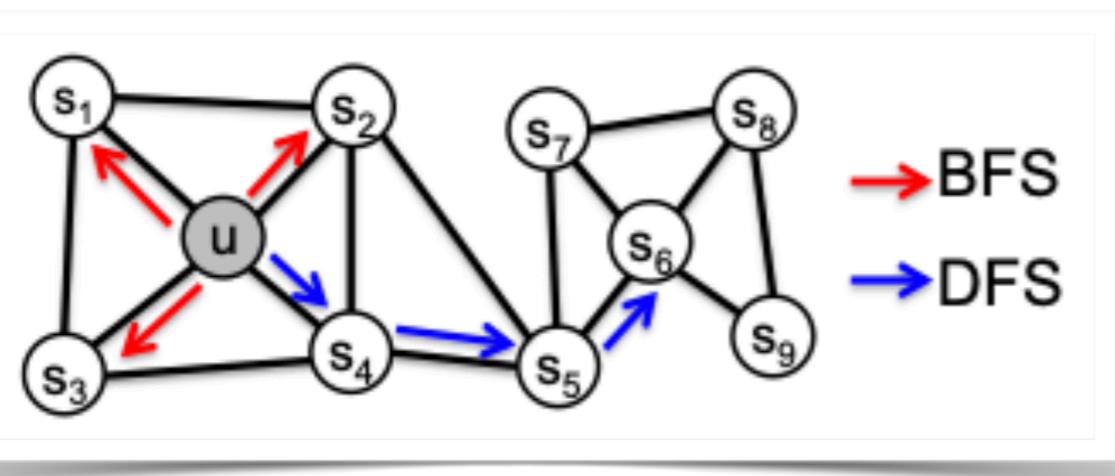
معرفی مسئله: انواع روش های تعیین شباهت نود (Node Similarity)

★ استراتژی های پیاده روی تصادفی



: از هر نود پیاده روی های تصادفی **Deepwalk** با طول ثابت و unbiased را اجرا می کنیم.

: مانند روشن deepwalk با تعریف دو پارامتر به منظور biased کردن پیاده روی های تصادفی (این دو پارامتر باعث یک توازن بین دید های محلی و جهانی شبکه می شود).



- پارامتر بازگشت
- پارامتر دور شدن

معرفی مسئله: انواع روش های تعیین شباht نود (Node Similarity)

★ استراتژی های پیاده روی تصادفی

- این روش علاوه بر مقیاس پذیری بسیار زیاد آن، برای شبکه های وزن دار، بی وزن، جهت دار و بی جهت به کار می رود.

★ اشکال این روش ها

- تنها برای شبکه های همگن خوب عمل می کنند.

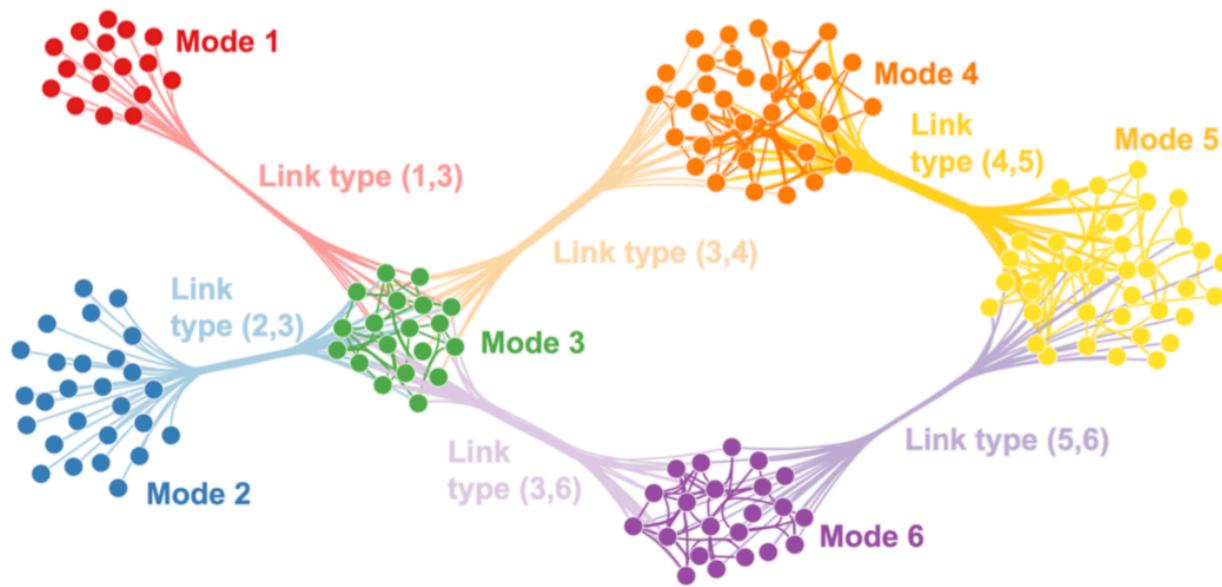
معرفی مسئله: انواع روش های تعیین شباخت نود (Node Similarity)

★ استراتژی های پیاده روی تصادفی

- این روش علاوه بر مقیاس پذیری بسیار زیاد آن، برای شبکه های وزن دار، بی وزن، جهت دار و بی جهت به کار می رود.

★ اشکال این روش ها

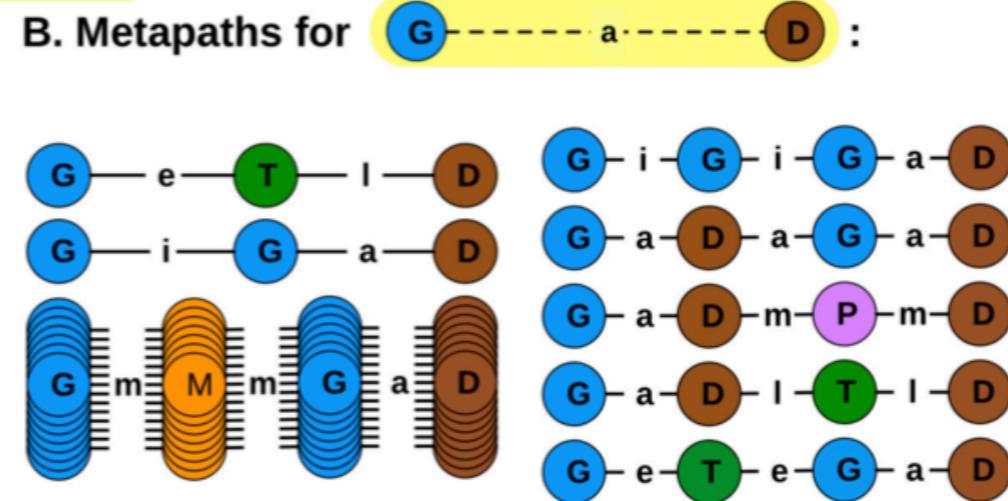
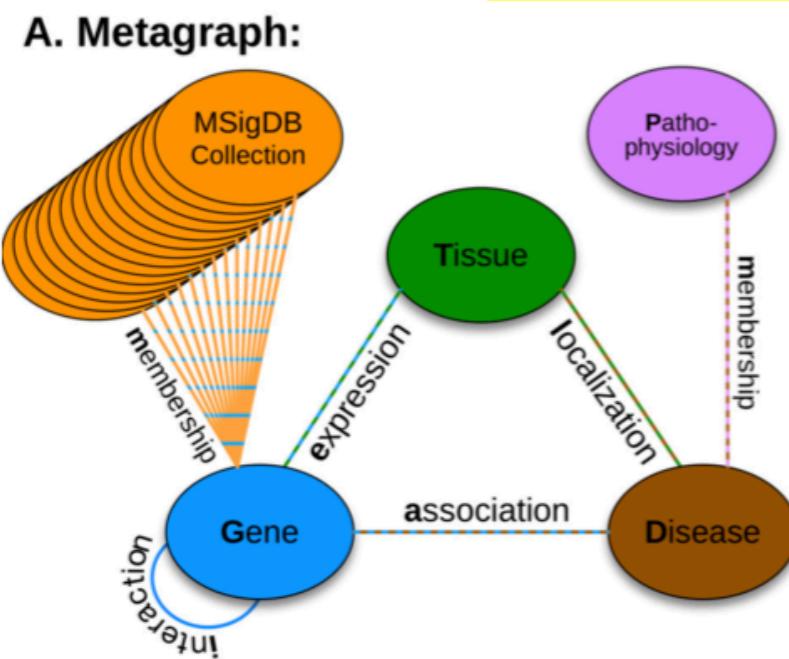
- تنها برای شبکه های همگن خوب عمل می کنند.



معرفی مسئله: انواع روش های تعیین شباخت نود (Node Similarity)

★ استراتژی های پیاده روی تصادفی

- این روش برای گراف های ناهمگن کاربرد دارد. در این روش، یک سری metapath تعریف می شود.



در این روش، همبستگی بین انواع مختلف نود را می گیرد.

معرفی مسئله: انواع روش های تعیین شباht نود (Node Similarity)

★ استراتژی های پیاده روی تصادفی

• معایب روش **metapath2vec**

برای تعریف metapath ها به دانش در آن حوزه نیازمندیم. ♦

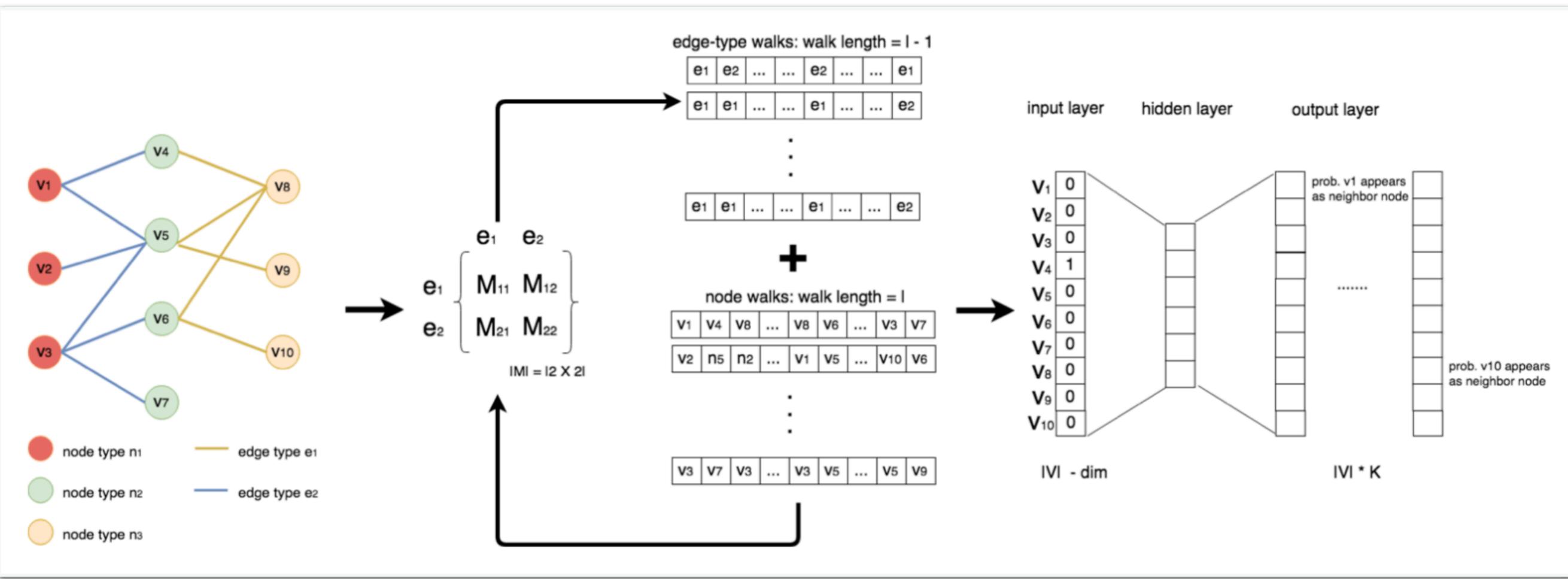
این روش، نوع یال ها را در نظر نمی گیرد. ♦

تنها از یک metapath در هر لحظه برای تولید یک پیاده روی تصادفی استفاده میکند. ♦

معرفی مسئله: انواع روش های تعیین شباht نود (Node Similarity)

★ استراتژی های پیاده روی تصادفی

: در این روش علاوه بر در نظرگرفتن مفاهیم نود، معانی یال ها را نیز در نظر می گیرد. **Edge2vec** • ✓

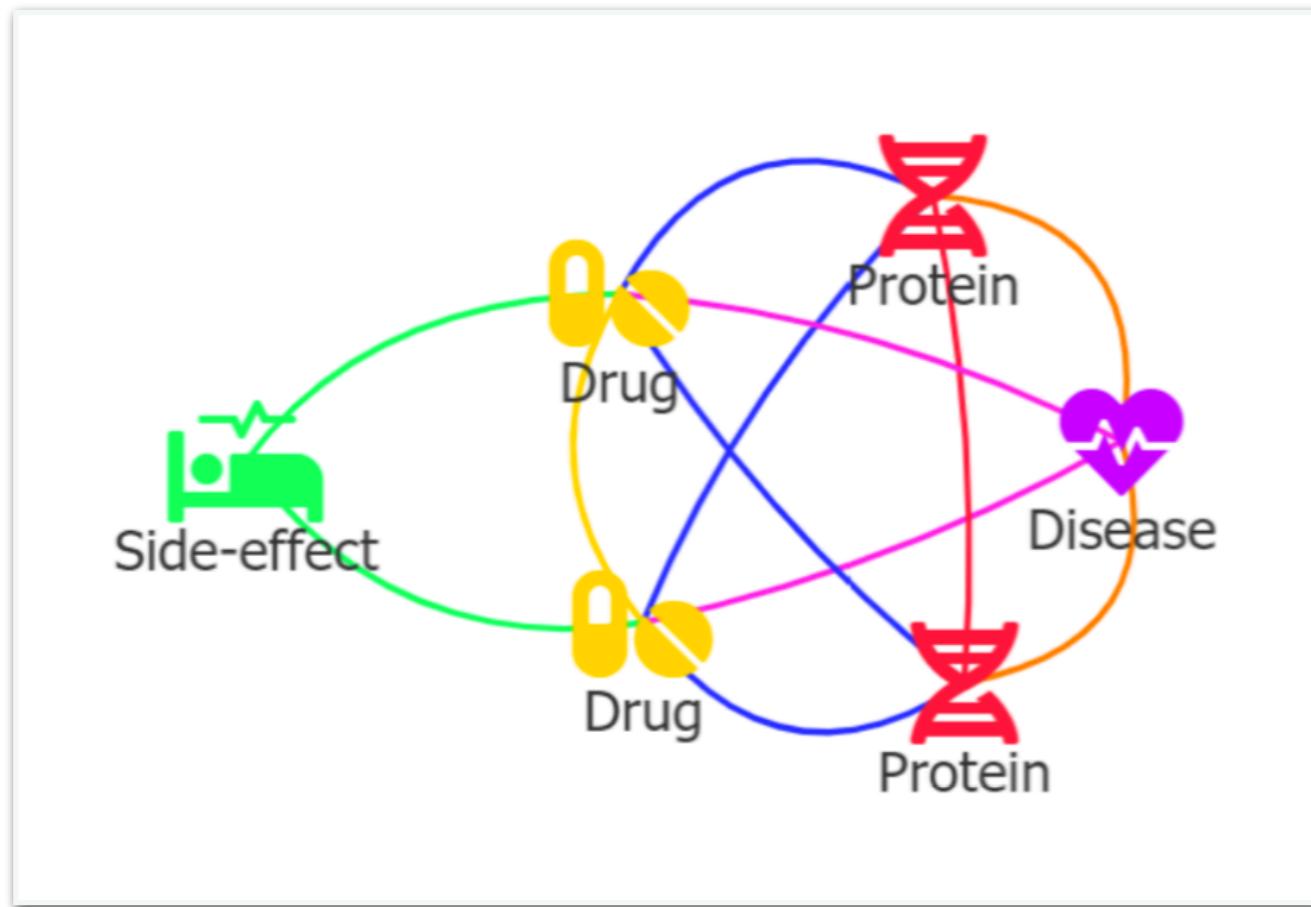


فهرست مطالب

- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

مجموعه داده مورد استفاده

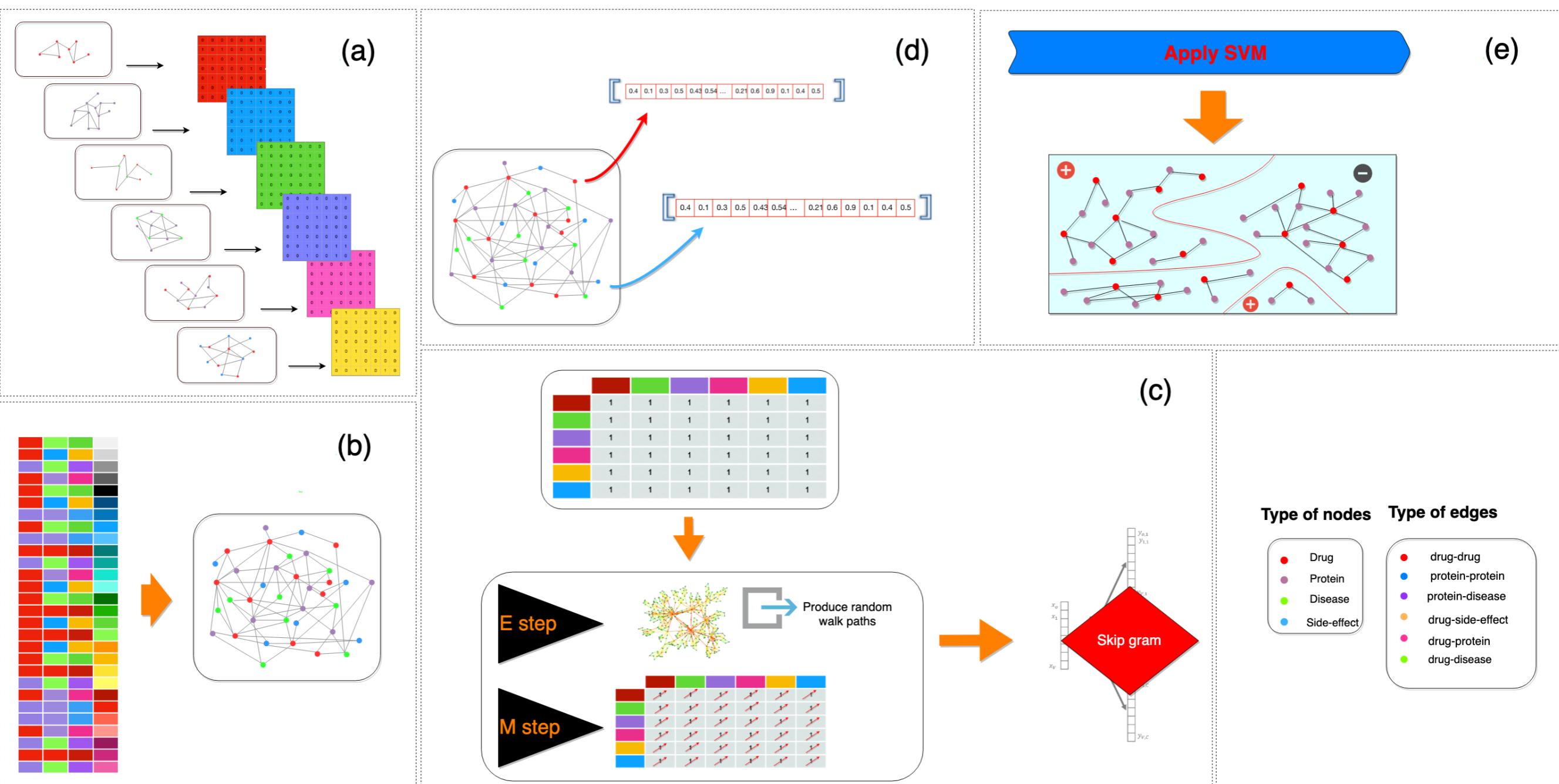
- تشکیل شده از چند پایگاه داده (Toxicogenomics و SIDER، HPRD، DrugBank)
- حاوی ۴ نوع نود (دارو، پروتئین، بیماری و اثرجانبی)
- متشکل از ۶ نوع یال (دارو-دارو، دارو-پروتئین، پروتئین-پروتئین، دارو-اثرجانبی، پروتئین-بیماری، دارو-بیماری)
- دارای ۱۲,۰۱۵ نود و ۱,۸۹۵,۴۴۵ یال



فهرست مطالب

- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

روش حل مسئله

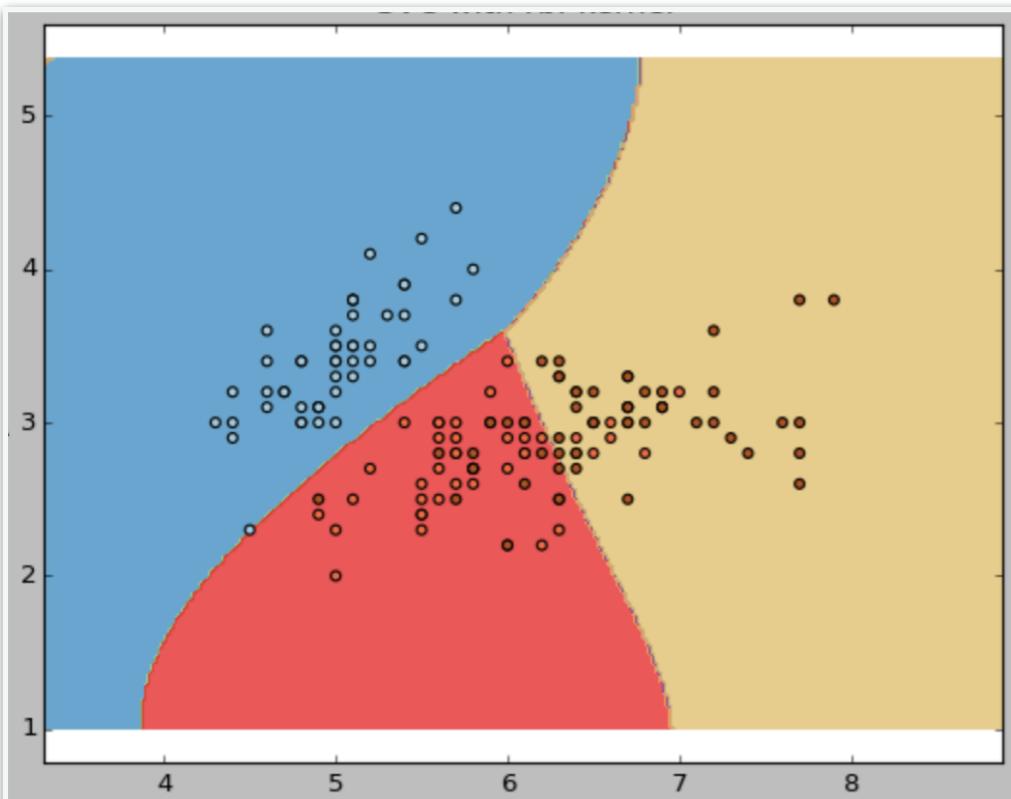


روش حل مسئله

بعد از تولید نمونه های منفی با دادن برچسب که نشان دهنده وجود ارتباط یا عدم وجود ارتباط می باشد، با استفاده از طبقه بند SVM روابط را پیش بینی نمودیم.

(SVM) Support Vector Machine

SVM یا ماشین بردار پشتیبان الگوریتم طبقه بندی یا classifier بوده و به عنوان یکی از بهترین تکنیک های دسته بندی و پیش بینی و تشخیص outlier شناخته می شود و در دسته یادگیری با ناظارت محسوب می شود.



- استفاده از کرنل RBF
- استفاده از Grid Search برای پیدا کردن بهترین مقادیر پارامترها

★ تنظیم پارامتر های edge2vec

• پارامتر های edge2vec

r, w, N, p •

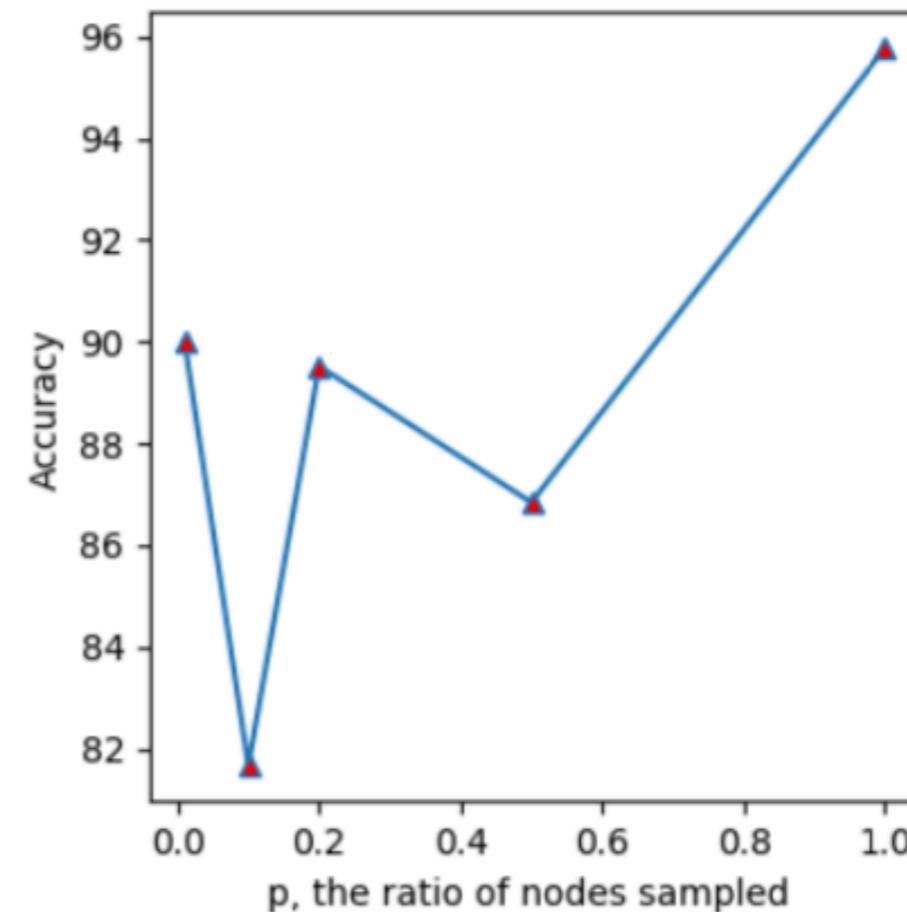
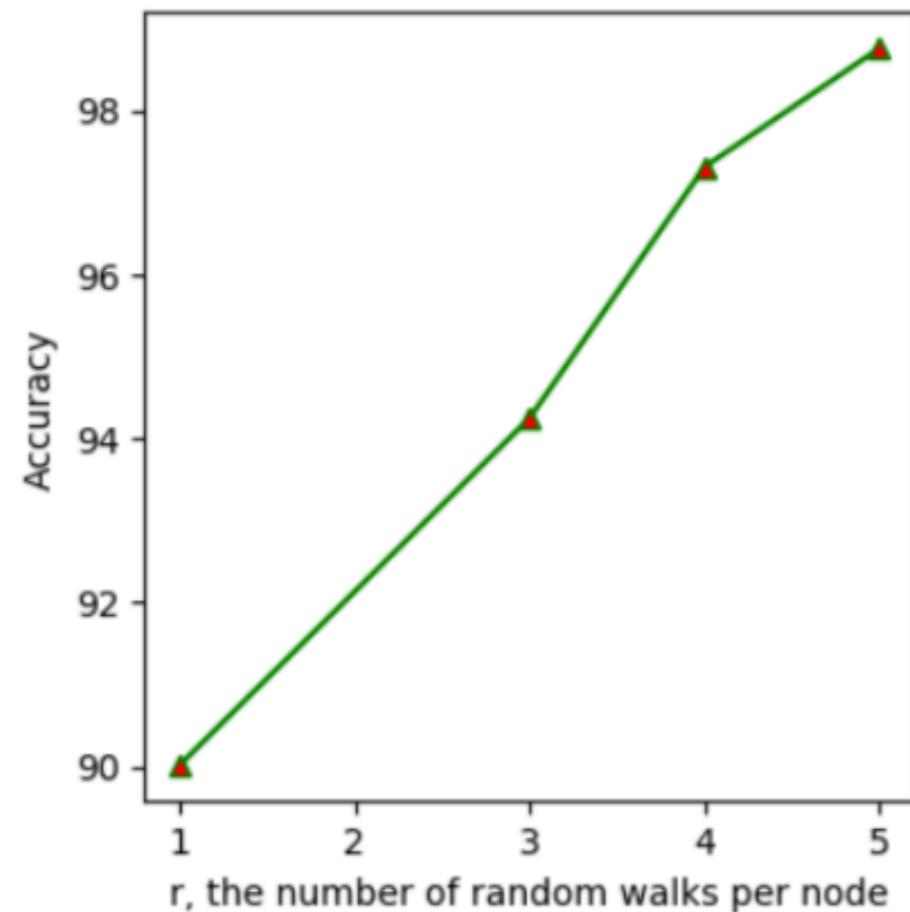
✓ تعداد walk‌ها روی هر نود $\leftarrow r$

طول walk در هر پیاده روی تصادفی $\leftarrow w$

تعداد تکرار برای آموزش ماتریس انتقال نوع لبه $\leftarrow N$

✓ نسبت نودهای نمونه برداری شده در فرایند آموزش ماتریس انتقال نوع لبه $\leftarrow P$

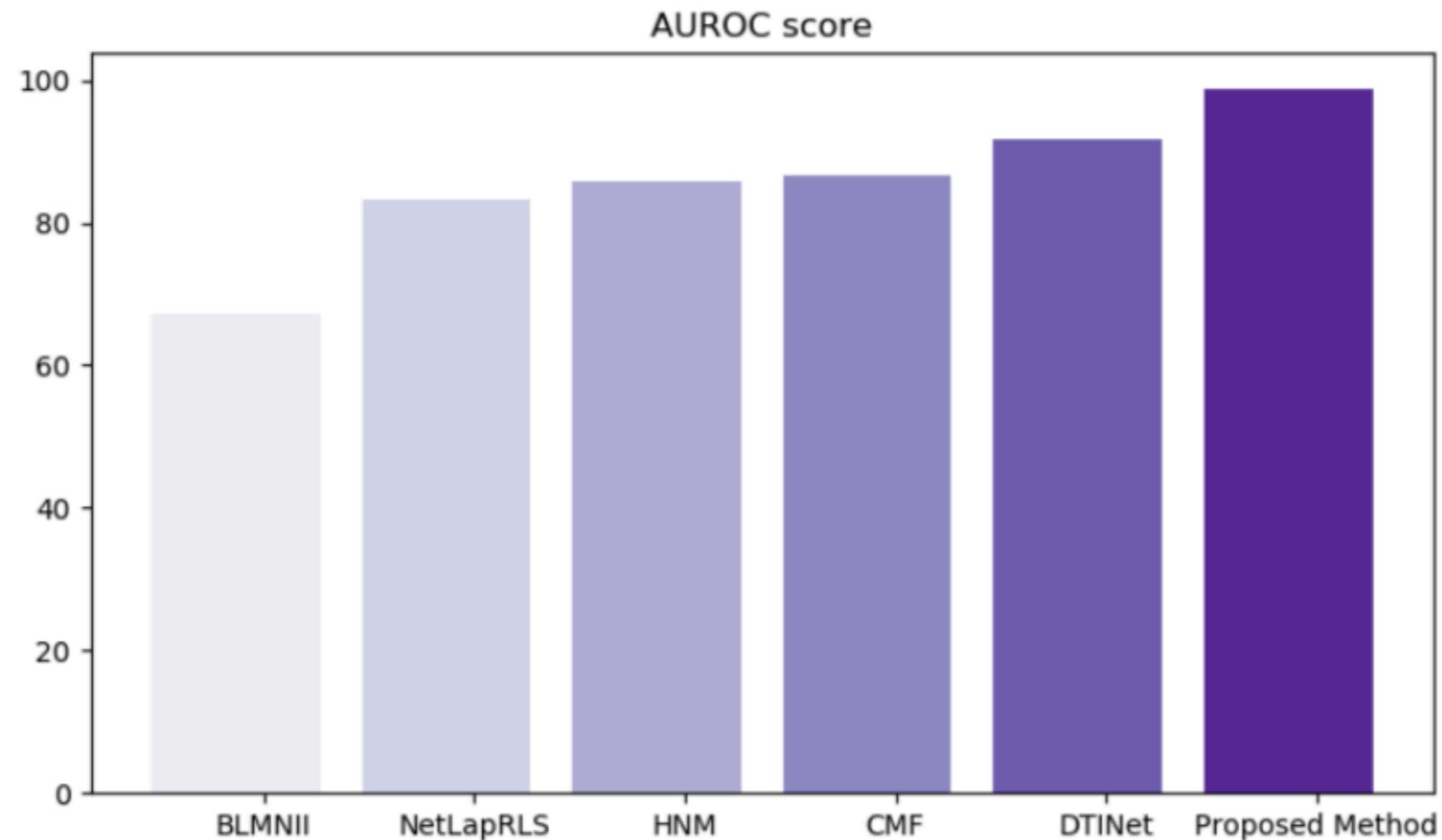
★ تنظیم پارامتر های edge2vec



فهرست مطالب

- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

		Recall	Precision	F1 score	Hamming loss
P	0.1	0.81	0.81	0.81	0.18
	0.2	0.89	0.89	0.89	0.10
	0.5	0.86	0.86	0.86	0.13
	1	0.95	0.95	0.95	0.04
r	3	0.94	0.94	0.94	0.05
	4	0.97	0.97	0.97	0.02
	5	0.98	0.98	0.98	0.01



فهرست مطالب

- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

کارهای آینده

- پیش بینی روابط بیشتری مانند DDI و PPI
- اجرای روش پیشنهادی بر روی مجموعه داده های دیگر

فهرست مطالب

- معرفی مسئله
- مقدمه ★
- شبکه (network) ★
- جاسازی نودها در گراف (node embedding) ★
- انواع روش های تعیین شباهت نود (Node Similarity) ★
- مجموعه داده مورد استفاده
- روش حل مسئله
- نتایج
- کارهای آینده
- مراجع

- J. A. DiMasi, “New drug development in the united states from 1963 to 1999,” *Clinical Pharmacology & Therapeutics*, vol. 69, no. 5, pp. 286–296, 2001.
- B. Perozzi, R. Al-Rfou, and S. Skiena, “Deepwalk: Online learning of social representations,” in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.
- J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, “Line: Large-scale information network embedding,” in *Proceedings of the 24th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1067–1077.
- A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.
- Y. Dong, N. V. Chawla, and A. Swami, “metapath2vec: Scalable representation learning for heterogeneous networks,” in *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2017, pp. 135–144.
- Z. Gao, G. Fu, C. Ouyang, S. Tsutsui, X. Liu, J. Yang, C. Gessner, B. Foote, D. Wild, Y. Ding, and Q. Yu, “edge2vec: Representation learning using edge semantics for biomedical knowledge discovery,” *BMC Bioinformatics*, vol. 20, no. 1, p. 306, Jun 2019. [Online]. Available: <https://doi.org/10.1186/s12859-019-2914-2>
- Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information,” *Nature communications*, vol. 8, no. 1, p. 573, 2017.

- C.Knox,V.Law,T.Jewison,P.Liu,S.Ly,A.Frolkis,A.Pon,K.Banco, C. Mak, V. Neveu *et al.*, “Drugbank 3.0: a comprehensive resource for ‘omics’ research on drugs,” *Nucleic acids research*, vol. 39, no. suppl_1, pp. D1035–D1041, 2010.
- T.KeshavaPrasad,R.Goel,K.Kandasamy,S.Keerthikumar,S.Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal *et al.*, “Human protein reference database—2009 update,” *Nucleic acids research*, vol. 37, no. suppl_1, pp. D767–D772, 2008.
- M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, “A side effect resource to capture phenotypic effects of drugs,” *Molecular systems biology*, vol. 6, no. 1, 2010.
- A. P. Davis, C. G. Murphy, R. Johnson, J. M. Lay, K. Lennon- Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, M. C. Rosenstein, T. C. Wiegers *et al.*, “The comparative toxicogenomics database: update 2013,” *Nucleic acids research*, vol. 41, no. D1, pp. D1104–D1114, 2012.
- W. Wang, S. Yang, X. Zhang, and J. Li, “Drug repositioning by integrating target information through a heterogeneous network model,” *Bioinformatics*, vol. 30, no. 20, pp. 2923–2930, 2014.
- X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, “Collaborative matrix factorization with multiple similarities for predicting drug-target interactions,” in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1025–1033.

با تشکر از توجه شما