Archive of SID

چکیده مبسوط چهارمین سمینار آنالیز تابعی و کابردهای آن

۱۳-۱۲اسفند ۱۳۹۴، دانشگاه فردوسی مشهد

کاربرد میانگین های ماتریسی در تشخیص ساختار انجمن های شبکه های اجتماعی

سمیرا اخوان [1] *، حامد وحدت نژاد [2]، و رمضان هاونگی[3]

[1] گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی بیرجند
akhavan.tak@gmail.com
[2] گروه مهندسی کامپیوتر، دانشگاه بیرجند
Vahdat.nejad@gmail.com
[3] گروه برق الکترونیک، دانشگاه بیرجند
Havangi@Birjand.ac.ir

چکیده. در این مقاله، با استفاده از مفهوم میانگین های ماتریسی و ماتریس های لاپلاس، الگوریتمی برای تشخیص ساختار انجمن ها در شبکه های اجتماعی ارائه می دهیم. الگوریتم داده شده برای گروه بندی داده ها به وسیله روش خوشه بندی طیفی بکار برده شده و برروی مجموعه داده استاندارد شبکه دلفین ها با دو روش فازی c-میانگین و k-میانگین اعمال شده است. در نهایت کیفیت الگوریتم ذکر شده به کمک نتایج بدست آمده با استفاده از معیار ماژلاریتی سنجیده شده است.

۱. پیش‌گفتار

یک انجمن در یک شبکه اجتماعی، از تجمع تعدادی عوامل و روابط میان آن ها تشکیل می شود که اعضای انجمن دارای بیشترین پیوند درون انجمن و کمترین پیوند بیرون آن هستند. در نظریه گراف ها، هر عامل یک گره و هر رابطه میان آنها یک یال است. این گره ها به وسیله انواع خاصی از وابستگی های متقابل مانند دوستی، خویشاندی، باورها، منافع مشترک و غیره به هم متصل شده اند.

تشخیص ساختار انجمن، به این دلیل که آنها بستری برای انتشار اطلاعات هستند، یک مساله مهم و چالش برانگیز است. یکی از الگوریتم های شناخته شده برای تشخیص انجمن ها، روش نیومن-گیروان [۱] است که در آن برای یافتن یال های انجمنی از مفهوم مرکزیت ارتباط استفاده می کند. الگوریتم با وجود بهبودهای حاصل شده دارای پیچیدگی زمان زیادی می باشد. الگوریتم دیگر برای شناسانی انجمن ها توسط نیومن [۴] پیشنهاد شد. این الگوریتم برای بیشینه سازی ماژولاریتی با قراردادن هر رأس در یک انجمن مجزا کار خود را شروع می کند و در شروع هیچ یالی وجود ندارد. با افزودن یک به یک یال ها انجمن هایی که در دوسر این یال ها قرار دارند

۲۰۱۰ Mathematics Subject Classification. Primary ۱۵A۱۸؛ Secondary ۶۵F۳۰.
واژگان کلیدی. میانگین ماتریسی، ماتریس لاپلاس، الگوریتم خوشه بندی طیفی .
* سخنران

www.SID.ir

Archive of SID

در صورت افزایش ماژولاریتی باهم ادغام می شوند. روش دیگر در این راستا الگوریتم "عابر تصادفی"است که توسط هوگز معرفی شد. در این روش از هر رأس با توجه به یال های موجود به هر کدام از رأس های مجاور بصورت تصادفی حرکت می شود. یک روش دیگر الگوریتم "گسترش برچسب ها" است. در این روش، رأس ها در ابتدا با برچسب های منحصر بفردی از هم مجزا شده و سپس در هر مرحله یکبار مورد بررسی قرار می گیرند و برچسب شان با کمک برچسب رأس های اطراف مشخص می شود؛ بدین ترتیب که هر رأس برچسب خود را به برچسبی که بیشترین تعداد تکرار را در بین رأس های همسایه دارد، تغییر می دهد.

روش مورد توجه در این مقاله، الگوریتم خوشه بندی طیفی [۳، ۵] است که مبنای آن استفاده از بردارهای ویژه ماتریس هایی خاص می باشد. در بخش بعد به توصیف این ماتریس ها می پردازیم.

۲. ماتریس لاپلاس

اساس روش خوشه بندی طیفی بر مبنای استفاده از ماتریس هایی موسوم به ماتریس لاپلاس است. معمولاً شبکه های اجتماعی را می توان به وسیله یک گراف نمایش داد.

تعریف ۱.۲. *فرض کنید $G = (V, E)$ یک گراف وزن دار و بدون جهت با مجموعه رأس های $V = \{v_1, \cdots, v_n\}$ باشد. درجه هر رأس v_i بصورت $d_i = \sum_{j=1}^{n} w_{ij}$ تعریف می شود که در آن w_{ij} وزن یال بین دو رأس v_i و v_j می باشد.*

فرض کنیم D ماتریس درجه و W ماتریس وزن های گراف باشد.

تعریف ۲.۲. *ماتریس های لاپلاس به سه نوع تقسیم می شوند:*

۱. *ماتریس لاپلاس غیر نرمال:*

$$L = D - W$$

۲. *ماتریس لاپلاس نرمال L_{rw}:*

$$L_{rw} = D^{-1}L = I - D^{-1}W$$

۳. *ماتریس لاپلاس نرمال L_{sym}:*

$$L_{sym} = D^{\frac{-1}{2}} L D^{\frac{-1}{2}}$$

روش های فراوانی برای معرفی شباهت بین داده ها وجود دارند. باید توجه کرد که چگونگی تعریف تشابه بین داده ها به طور مستقیم روی ساختار انجمن های به دست آمده تاثیر می گذارد. در این مقاله، تابع شباهت را براساس همسایگان مشترک معرفی می کنیم، چرا که در شبکه های اجتماعی افرادی که دوست مشترک بسیاری دارند به احتمال زیاد با یکدیگر دوست بوده و به گروه مشترکی تعلق دارند. بنابراین تابع شباهت برای الگوریتم پیشنهادی بدین صورت تعریف می کنیم.

تعریف ۳.۲. *فرض کنید k عددی صحیح و مثبت باشد. همسایگی k-فاصله برای رأس v شامل تمام رأس هایی است که فاصله آنها از v بزرگتر از k نباشد. به عبارت دیگر*

$$N_k(v) = \{w \in V;\ \ d(w, v) \leq k\}\,.$$

www.SID.ir

Archive of SID

بنابراین $N_1(v)$ شامل همه رأس هایی است که با v مجاور هستند.

تعریف ۴.۲. تابع شباهت را به صورت زیر تعریف می‌کنیم:

$$S(w,v)=\begin{cases}\frac{|N_k(v)\cap N_k(w)|+1}{\sqrt{|N_k(v)|.|N_k(w)|}}, & v,w\in E,\ v\neq w\\ 0 & v,w\notin E,\ v\neq w\\ 1 & v=w\end{cases}$$

۳. الگوریتم پیشنهادی

فرض می‌کنیم که مجموعه داده مورد نظر شامل n نقطه $x_1,\cdots,x_n$ است. تشابه بین جفت داده x_i و x_j را با استفاده از تابع شباهت تعریف ۴.۲ سنجیده و بصورت $s_{ij}=S(x_i,x_j)$ نمایش داده می شود. ماتریس به دست آمده توسط اعداد s_{ij} که یک ماتریس مثبت و متقارن است، ماتریس شباهت نامیده می شود.

انتخاب ماتریس لاپلاس مورد استفاده در خوشه بندی طیفی از اهمیت زیادی برخوردار است بطوری که اگر گراف منظم باشد و اکثر راس های آن دارای درجه تقریبا یکسانی باشند، آنگاه تمام ماتریس های لاپلاس مشابه یکدیگر هستند و می توان از هرکدام برای خوشه بندی استفاده کرد. اما اگر درجه رأس های گراف تفاوت زیادی با یکدیگر داشته باشند، ماتریس های لاپلاس بطور قابل توجهی متفاوت خواهند بود. استدلال های زیادی وجود دارند که تضمین می‌کند استفاده از ماتریس های لاپلاس نرمال به مراتب بهتر از ماتریس لاپلاس غیرنرمال می باشد.

اما با استفاده از مفهوم میانگین ماتریس ها، ماتریس لاپلاس جدیدی را بصورت زیر تعریف می‌کنیم:

$$L_\sharp=L_{rw}^{\frac{1}{2}}\left(L_{rw}^{\frac{-1}{2}}L_{sym}L_{rw}^{\frac{-1}{2}}\right)^{\frac{1}{2}}L_{rw}^{\frac{1}{2}}. \qquad (۱.۳)$$

بطور خلاصه مراحل الگوریتم به شرح زیر است:

ورودی: گراف شبکه اجتماعی، تعداد خوشه های مورد نظر

خروجی: ساختار شبکه تقسیم شده به خوشه ها

گام ۱: ساخت گراف شباهت با استفاده از تعریف ۴.۲ و قراردادن آن بعنوان ماتریس وزن .

گام ۲: محاسبه میانگین لاپلاس های نرمال از طریق معادله (۱.۳).

گام ۳: محاسبه k-بردار ویژه متناظر با k-مقدار ویژه کوچکتر ماتریس لاپلاس.

گام ۴: محاسبه ماتریس $U\in\mathbb{R}^{n\times k}$ که ستون های آن را بردارهای ویژه تشکیل می دهند.

گام ۵: مشخص کردن بردارهای برای هر ، که همان سطرهای ماتریس هستند.

گام ۶: خوشه بندی نقاط داده با استفاده از الگوریتم فازی c-میانگین (k-میانگین) به دسته های $C_1,\ldots,C_k$.

www.SID.ir

Archive of SID

۴. ارزیابی

الگوریتم داده شده روی مجموعه داده استاندارد شبکه دلفین ها [۲] اعمال شده است. در جداول زیر مقدار ماژولاریتی[۴] ساختار انجمن های به دست آمده از طریق الگوریتم های فازی c-میانگین و k-میانگین در مورد ماتریس های لاپلاس آمده است. مقدار ماژولاریتی (که یک معیار استاندارد ارزیابی کیفیت خوشه بندی است) ساختار انجمن با استفاده از $L_\sharp$ بطور قابل توجهی بهبود پیدا کرده است.

	L_{sym}	L_{rw}	$L_\sharp$
الگوریتم c-میانگین	۰/۵۶۹۹	۰/۶۰۵۵	۰/۶۲۰۳
الگوریتم k-میانگین	۰/۴۲۶۶	۰/۵۸۵۳	۰/۶۴۸۸

مراجع

۱. M. Girvan، M.E.J. Newman، *Community structure in social and biological networks*، Proceedings of the National Academy of Sciences، ۹۹ ،(۲۰۰۲) ۷۸۲۶–۷۸۲۱.

۲. D. Lusseau، *The emergent properties of a dolphin social network*، Proceedings of the Royal Society of London، Series B، ۲۷۰ ،(۲۰۰۳) S۱۸۶–S۱۸۸.

۳. U.v. Luxburg، *A tutorial on spectral clustering*، Statistics and Computing، ۱۷ ،(۲۰۰۷) ۴۱۶–۳۹۵.

۴. M.E.J. Newman، *Fast algorithm for detecting community structure in networks*، Physical review E، ۶۹ ،(۲۰۰۴) ۰۶۶۱۳۳.

۵. P. G. Sun، L.Gao and S. S. Han، *Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks*، Information Sciences، ۱۸۱ ،(۲۰۱۱) ۱۰۷۱–۱۰۶۰.

www.SID.ir