

مقدار دهی اولیه گرادیان مزدوج در خوشه‌بندی طیفی با الگوریتم ژنتیک

مهدی نعمتی، محمود امین‌طوسی، مهدی زعفرانیه^۱

دانشکده ریاضی و علوم کامپیوتر، دانشگاه حکیم سبزواری، سبزوار، ایران

چکیده

خوشه‌بندی داده‌ها از جمله مسائل مهم حوزه داده‌کاوی و یادگیری ماشین است که کاربردهای فراوانی در علوم مختلف دارد. یکی از روش‌های خوشه‌بندی، نمایش وابستگی داده‌ها توسط گراف و استفاده از شیوه‌های افرازبندی گراف می‌باشد. خوشه‌بندی طیفی گراف یکی از مشهورترین این شیوه‌هاست که توجه زیادی را به خود معطوف نموده است. در روش خوشه‌بندی طیفی گراف از بردارویژه متناظر با دومین کوچکترین مقدار ویژه برای خوشه‌بندی داده‌ها استفاده می‌شود. به جای محاسبه مقادیر ویژه و بردارهای ویژه، می‌توان از الگوریتم‌های بهینه‌سازی همچون گرادیان مزدوج نیز برای پیدا کردن پاسخ مسئله استفاده نمود. شیوهی گرادیان مزدوج، روشی تکراری است که از یک نقطه آغازین تصادفی حرکت نموده و به سمت جواب کمینه حرکت می‌کند. بسته به نقطه شروع اولیه، الگوریتم ممکن است یک کمینه محلی را بیابد. در این مقاله با استفاده از الگوریتم ژنتیک، نقطه شروع مناسبی برای این روش خوشه‌بندی پیدا می‌شود. نتایج آزمایشات کارایی شیوه پیشنهادی را نشان داده است.

واژه‌های کلیدی: خوشه‌بندی طیفی گراف، بردارویژه، مقدارویژه، گرادیان مزدوج، الگوریتم ژنتیک

Mathematics Subject Classification [2010]: 91C20, 97K30.

۱ مقدمه

گراف‌ها ساختارهای ریاضی هستند که می‌توان از آن‌ها برای مدل‌سازی دامنه وسیعی از مسائل استفاده کرد. بسیاری از موارد کاربردی، نیاز به خوشه‌بندی گراف‌ها و افراز آن‌ها به زیرگراف‌ها دارند. در این میان خوشه‌بندی طیفی گراف، استفاده از مقادیر ویژه و بردارهای ویژه ماتریس لاپلاسین متناظر با گراف را برای خوشه‌بندی ارائه می‌کند. به دنبال آن، روش گرادیان مزدوج به دلیل سرعت در همگرایی و میزان حافظه مورد نیاز کم، برای تقریب زدن این بردارها مورد توجه است. در این روش با شروع از یک نقطه اولیه، مقدار تابع هدف کمینه می‌شود. گاهی این روش به سمت نقاط مینیم محلی همگرا می‌شود که در شرایط مساله صدق می‌کنند اما افرازبندی حاصل از آن مطلوب نیست. در این مقاله به کمک الگوریتم ژنتیک، جایگزینی از مولفه‌های نقطه اولیه را می‌یابیم که الگوریتم گرادیان مزدوج به سمت مینیم سراسری همگرا شود و در نتیجه افرازبندی مطلوب را نتیجه دهد.

گراف ساده و بدون جهت $G = (V, E)$ را در نظر بگیرید که $V = \{1, 2, \dots, n\}$ مجموعه رئوس گراف بوده به گونه‌ای که هر گره با یک عدد واحد $i \in V$ نمایش داده می‌شود و اندازه گراف برابر با $|V| = n$ است. مجموعه یال‌های گراف $E = \{(i_1, j_1), \dots, (i_e, j_e)\}$ نشان‌دهنده e یال در گراف است که هر یال از i به j با زوج $(i, j) \in E$ نمایش داده می‌شود. هدف از خوشه‌بندی گراف، افراز گراف به دو زیرگراف با اندازه‌های برابر است به گونه‌ای که تا حد ممکن تعداد یال‌های متصل کننده دو زیرگراف کمینه شود.

تعریف ۱.۱. ماتریس مجاورت گراف G به صورت زیر تعریف می‌شود:

$$\begin{cases} W_{ij} = 1 & \text{if } (i, j) \in E \\ W_{ij} = 0 & \text{if } (i, j) \notin E \end{cases} \quad \forall i, j \in V$$

که $W_{n \times n}$ یک ماتریس متقارن است.

^۱سخنران

ماتریس درجه D ، ماتریسی قطری است که d_{ii} درجه رأس i ام در گراف G است. ماتریس لاپلاسین گراف از فرمول $L = D - W$ به دست می‌آید که متقارن و معین مثبت است [۳].

۲.۱. گراف ساده و بدون جهت $G = (V, E)$ به صورتی که $|V| = n$ را در نظر بگیرید، فرض کنید $X \in \{-1, 1\}^n$ باشد. گراف را به دو زیرگراف افراز می‌کنیم و برای گره‌هایی که به افراز اول نسبت داده شده‌اند x_i متناظر با آن را برابر $1+$ و در غیر این صورت، -1 در نظر می‌گیریم. همچنین، فرض کنید L ماتریس لاپلاسین متناظر با گراف G باشد. در این صورت، تعداد یال‌های بین دو زیرگراف که توسط X تولید شده‌اند، برابر است با $\frac{1}{4} (X^t L X)$ [۲].

با استفاده از ل ۲.۱، با انجام یک آزادسازی و افزودن محدودیت $X^t X = n$ به تابع $f(X) = \frac{1}{4} (X^t L X)$ ، محدودیت عدد صحیح بودن X برداشته می‌شود. همچنین برای برابر بودن اندازه دو زیرگراف، محدودیت $e^t X = 0$ به مساله اضافه و در نهایت، مساله بهینه‌سازی زیر حاصل می‌شود [۲]:

$$\begin{aligned} \min \quad & f(x) = X^t L X \\ \text{s.t.} \quad & e^t X = 0 \\ & \sum_{i=1}^n x_i^2 = n \end{aligned} \quad (1)$$

۳.۱. فرض کنید $A \in M_n$ یک ماتریس هرمیتی باشد، و $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ مقادیر ویژه A باشند، و $i_1, \dots, i_k$ اعداد صحیح باشند به طوری که $1 \leq i_1 < \dots < i_k \leq n$ ، و $x_{i_1}, \dots, x_{i_k}$ متعامد باشند به گونه‌ای که $Ax_{i_p} = \lambda_{i_p} x_{i_p}$ برای هر $p = 1, \dots, k$ ، و همچنین $S = \text{span}\{x_{i_1}, \dots, x_{i_k}\}$ ، آنگاه

$$\begin{aligned} \lambda_{i_1} &= \min_{\{x: x \neq 0, x \in S\}} \frac{x^t A x}{x^t x} = \min_{\{x: x \in S \text{ and } \|x\|_2=1\}} x^t A x \\ &\leq \max_{\{x: x \in S \text{ and } \|x\|_2=1\}} x^t A x = \max_{\{x: x \neq 0, x \in S\}} \frac{x^t A x}{x^t x} = \lambda_{i_k} \end{aligned} \quad (2)$$

اگر $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ نشان‌دهنده مقادیر ویژه ماتریس $M \in \mathbb{R}^{n \times n}$ باشند، آنگاه، $\lambda_k = \min_k \frac{X^t M X}{X^t X}$ که بردار ویژه متناظر با مقدار ویژه λ_k جواب آن است [۱].

از ل ۳.۱ نتیجه می‌شود λ_1 کمترین مقدار این تقسیم و سپس λ_2 مقدار کمینه بعدی و به ترتیب تا آخر، در نهایت λ_n بیشترین مقدار این تقسیم است. تقسیم $X^t M X / X^t X$ به عنوان خارج قسمت رایلی یا تقسیم ریلی-ریتر^۲ شناخته می‌شود. برای ماتریس لاپلاسین گراف، $\lambda_1 = 0$ و بردار ویژه متناظر با آن $e = (1, \dots, 1)^t$ است. در نتیجه بردار ویژه متناظر با کوچک‌ترین مقدار ویژه، جواب مساله (۱) است، اما در محدودیت $e^t X = 0$ صدق نمی‌کند، از این رو از مینیمم دوم، یعنی دومین کوچک‌ترین مقدار ویژه استفاده می‌شود و در نهایت به تابع هدف زیر می‌رسیم [۴].

$$f(X) = \frac{X^t L X}{X^t X}$$

۱.۱ الگوریتم گرادیان مزدوج

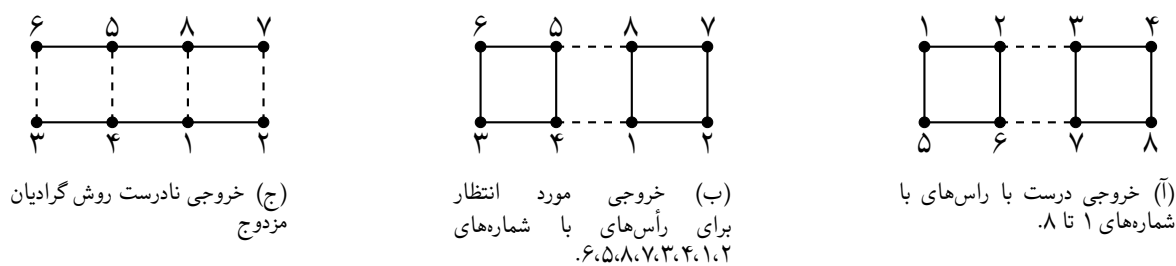
روش گرادیان مزدوج، یک الگوریتم تکراری است که از یک نقطه اولیه X شروع می‌کند و در n گام (n بعد مساله است) به نقطه مینیمم محلی همگرا می‌شود. محاسبه نقاط در این روش به صورت زیر است [۲]:

$$X_k = X_{k-1} + \alpha_k d_k \quad (3)$$

که در آن

$$\alpha_k = \arg \min_{\alpha > 0} f(X_k + \alpha d_k)$$

^۲Rayleigh quotient



شکل ۱: خط‌چین‌ها معرف لبه‌های مورد برش در گراف هستند.

همچنین در رابطه (۳)، d_k جهتی است که در راستای آن مقدار تابع هدف کاهش می‌یابد و به روش زیر به دست می‌آید:

$$d_0 = -g_0, \quad d_k = -g_{k-1} + \beta_{k-1} d_{k-1} \quad (4)$$

که در آن g_k گرادیان تابع f در نقطه x_k و β_{k-1} پارامتر به‌هنگام‌سازی جهت‌های مزدوج است

۲ روش پیشنهادی: مقداردهی نقطه اولیه روش گرادیان مزدوج با الگوریتم ژنتیک

بردارهای ویژه ماتریس لاپلاسین، که جواب مورد نظر ما هم جزو آن‌هاست، با یکدیگر متعامدند [۳]. در نتیجه ضرب داخلی آن‌ها برابر صفر می‌باشد. با انتخاب نقطه اولیه X_0 به گونه‌ای که $x_i = i - (|V| + 1)/2$ ($\forall i = 1, 2, \dots, |V| = n$)، نقطه اولیه X_0 و بردار ویژه e متعامد می‌شوند. به دلیل مزدوج بودن جهات در روش گرادیان مزدوج تمام نقاط بدست آمده بعدی نیز با بردار e متعامدند [۱].

به دلیل همگرایی الگوریتم گرادیان مزدوج به مینیمم محلی، انتخاب نقطه اولیه در رسیدن به بهینه سراسری حائز اهمیت است. همان گونه که در شکل ۱ دیده می‌شود، با اجرای الگوریتم گرادیان مزدوج بر روی ماتریس لاپلاسین این گراف، خروجی الگوریتم، شکل ۱ (ج) می‌شود که با چهار برش، گراف را به دو زیرگراف با رأس‌های $V_1 = \{1, 2, 3, 4\}$ و $V_2 = \{5, 6, 7, 8\}$ افزایش می‌کند، در حالی که خروجی بهینه که در شکل ۱ (ب) نشان داده شده است، با دو برش به دو زیرگراف با رأس‌های $V_1 = \{1, 2, 7, 8\}$ و $V_2 = \{3, 4, 5, 6\}$ می‌باشد.

اگر درایه‌های X_0 را جابجا کنیم، همچنان X_0 با بردار e متعامد باقی می‌ماند. می‌توان با جابجایی این درایه‌ها نقطه اولیه جدید را تولید کرد. ما به کمک الگوریتم ژنتیک به دنبال جایگشت مناسبی از این درایه‌ها هستیم تا این نقطه اولیه، الگوریتم گرادیان مزدوج را به سمتی هدایت کند که مقدار تابع هدف مساله (۱) تا حد ممکن مینیمم شده و کمترین تعداد یال برشی را نتیجه دهد.

۳ نتایج آزمایشات

در شکل ۱ به عنوان نمونه یک گراف کوچک نشان داده شد که در برخی اجراهای الگوریتم گرادیان مزدوج گراف مطابق شکل ۱ (ج) به اشتباه به دو بخش تقسیم می‌شود؛ در شکل ۱ (ج) لبه‌های نشان داده شده با خط چین، یال‌های برشی هستند. به منظور بررسی کارایی شیوه پیشنهادی در بهبود روش گرادیان مزدوج از طریق جستجوی نقطه شروع اولیه مناسب، چندین گراف تصادفی با ابعاد مختلف تولید شده و در آزمایشات بکار گرفته شده‌اند. نتایج آزمایشات در جدول‌های ۱ تا ۳ آمده است. شماره ردیف در هر جدول بیانگر شماره مورد استفاده است. گراف‌های مورد بررسی، گراف‌های تصادفی هستند که به ترتیب دارای $150, 200, 250, 300$ و 350 رأس می‌باشند. تابع هدف مورد استفاده در الگوریتم ژنتیک، تابع $X^t L X$ بوده است که نتایج مقایسه‌ای بر مبنای این تابع هدف در جدول ۱ آمده است. برای هر اندازه‌ی گراف، 20 گراف تصادفی تولید شده و میانگین نتایج در جداول آمده است. به این ترتیب به عنوان نمونه، سطر اول جدول فوق‌الذکر، نتایج اجرا بر روی 20 گراف تصادفی با 150 رأس می‌باشد. سطر آخر هر جدول میانگین هر ستون را نشان می‌دهد. همان‌گونه که در ۱ دیده می‌شود، الگوریتم ژنتیک، توانسته است در تمام مثال‌ها نتایج روش گرادیان مزدوج را بهبود بخشد.

با خوشه‌بندی بردار X حاصل از گرادیان مزدوج می‌توان بردار جدید شامل $+1$ و -1 را بدست آورد که با قرار دادن این بردار در رابطه $X^t L X$ از لم ۲.۱ تعداد یال‌های برشی قابل محاسبه است. گراف‌های تولید شده در آزمایشات انجام شده، گراف‌هایی بدون وزن و بدون جهت بوده‌اند. لذا در ماتریس مجاورتی گراف، وجود لبه با یک و عدم وجود لبه با صفر نشان داده شده است. از این رو، مجموع تعداد لبه‌های برشی بین دو خوشه، محکی برای میزان بهینه بودن خوشه‌بندی انجام شده می‌باشد. جدول ۲ تعداد لبه‌های برشی بین دو خوشه در دو روش مورد بررسی را نشان می‌دهد. همان‌گونه که انتظار می‌رود، این نتایج مؤید نتایج جدول قبلی هستند. مجدداً در اینجا هم نتایج هر سطر میانگین 20 اجرای مختلف هستند.

روش		شماره
GACG	CG	
۳/۱۷۳	۰/۰۸۰	۱
۲/۷۶۰	۰/۰۰۷	۲
۳/۲۲۲	۰/۰۰۹	۳
۳/۷۸۸	۰/۰۱۲	۴
۴/۳۹۱	۰/۰۱۳	۵
۳/۴۶۷	۰/۰۲۴	میانگین

جدول ۳: زمان اجرا

روش		شماره
GACG	CG	
۳۱۳/۰۰۰	۳۳۴/۲۰۰	۱
۴۵۹/۹۰۰	۴۹۰/۶۵۰	۲
۶۷۹/۵۰۰	۷۲۶/۴۰۰	۳
۹۵۷/۵۵۰	۹۹۹/۸۰۰	۴
۱۲۸۸/۶۰۰	۱۳۳۶/۹۵۰	۵
۷۳۹/۷۱۰	۷۷۷/۶۰۰	میانگین

جدول ۲: تعداد یال‌های برشی

روش		شماره
GACG	CG	
۷/۷۶۶	۸/۹۹۱	۱
۸/۷۸۵	۹/۸۶۳	۲
۱۰/۴۱۴	۱۱/۷۲۶	۳
۱۲/۱۴۳	۱۳/۲۴۸	۴
۱۴/۲۴۹	۱۵/۲۰۰	۵
۱۰/۶۷۲	۱۱/۸۰۵	میانگین

جدول ۱: هزینه تابع هدف

بهبود حاصله در نتیجه استفاده از الگوریتم ژنتیک برای پیدا کردن مقادیر اولیه مناسب روش گرادیان مزدوج، زمان پردازش را بیشتر کرده است. هر سطر جدول ۳ زمان موردنیاز برای تمام ۲۰ نمونه تصادفی از هر اندازه گراف برای دو روش را نشان می‌دهد. تمام آزمایشات در نرم‌افزار MATLAB 2017 و بر روی دستگاه دارای پردازشگر مرکزی Intel CORE i7 با ۸ هسته انجام شده است.

۴ نتیجه‌گیری

الگوریتم گرادیان مزدوج در مواردی نتایج مورد انتظار را نداشته و طبیعت الگوریتم به گونه‌ای است که در این موارد راهی برای مقایسه و بهبود نتیجه ندارد. در بیشتر موارد مشاهده شد که نقاط اولیه به دست آمده از الگوریتم ژنتیک، الگوریتم گرادیان مزدوج را به سمت مینیمم سراسری همگرا می‌کند که در نتیجه تعداد یال‌های برشی به کمترین مقدار می‌رسد و این مشکل را به خوبی حل می‌کند.

مراجع

- [۱] اس.اس. رائو، ترجمه شهیدی‌پور، سید محمد مهدی، بهینه‌سازی (تئوری و کاربردی)، انتشارات دانشگاه فردوسی مشهد، مشهد، ۱۳۷۳.
- [۲] احمد ابویی‌مه‌ریزی، رضا قنبری، هشتمین کنفرانس بین‌المللی تحقیق در عملیات، بررسی تاثیر عددی الگوریتم گرادیان مزدوج ترکیبی بابایی و قنبری در حل مساله افزایش‌بندی گراف (۱۳۹۴).
- [3] N.P. Krut, Parallel Computing, *A conjugate gradient method for the spectral partitioning of graphs* volume number (1997) 1493-1502.
- [4] A. Pothen, H.D. Simon and K.P. Liou, SIAM Journal on Matrix Analysis and Applications, *Partitioning sparse matrices with eigenvectors of graphs* volume number (1990) 430-452.

پست الکترونیکی: m.nemati@sun.hsu.ac.ir

پست الکترونیکی: m.amintoosi@hsu.ac.ir

پست الکترونیکی: mehdi.zaferanieh@gmail.com