انتخاب متغیر با استفاده از رگرسیون انقباضی و کاربرد آن در داده‌های ریزآرایه

زهرا جمشیدنژاد، آرش اردلان
گروه آمار، دانشگاه یاسوج

چکیده

در مطالعات رگرسیونی، زمانی که بین متغیرهای مستقل همبستگی بالایی وجود داشته باشد استفاده از روش‌های معمول از جمله روش کمترین مربعات معمولی باعث ناپایداری واریانس برآوردها می‌شود. یک راه حل معمول، استفاده از روش کمترین مربعات جریمه‌دار است که در آن برای مقادیر بزرگ برآوردگرها، جریمه بالایی در نظر گرفته می‌شود و به نوعی تغییرات برآوردگر تحت کنترل در می‌آید.

مورد دیگر استفاده از رگرسیون جریمه‌دار در مدل‌های با ابعاد بالا یعنی مدل‌هایی با تعداد زیادی متغیر مستقل است. در این مدل‌ها تلاش می‌شود از ضرایب " نزدیک به صفر " حتی‌الامکان صرف‌نظر گردد تا فقط متغیرهایی در مدل باقی بمانند که تأثیر کاملاً معنی‌داری در متغیر وابسته دارند.

در این مقاله تلاش شده است ضمن مرور مختصری بر روش کمترین مربعات جریمه‌دار، رگرسیون جریمه‌دار و نحوه عملکرد این روش در برازش مدل‌هایی با ابعاد بالا مورد مطالعه و بررسی قرار گیرد. با ارائه دو سری داده واقعی، درستی بعضی از روابط و برتری این روش در مقایسه با سایر روش‌ها تحقیق شده است و سپس از برآوردهای این روش در تحلیل داده‌ها مورد استفاده قرار گرفته است.

واژه‌های کلیدی: انتخاب متغیر، الاستیک‌نت، برآوردیابی، داده‌های با ابعاد بالا، ستیغی، شبکه مقید، لاسو، ماتریس لاپلاس.

۱ مقدمه

با پیشرفت علوم مختلف وانجام تحقیقات پیشرفته، هر روز انواع مختلفی از داده‌ها مطرح می‌شود که نیازمند روش‌های قابل اطمینان برای تجزیه و تحلیل هستند.
در دنیای واقعی عوامل متعددی ممکن است بر روی یک پدیده تأثیرگذار باشند. در علوم مختلفی همچون ژنتیک، پزشکی، زیست شناسی و..... معمولاً با انبوه عوامل موثر بر یک پدیده که معمولاً این عوامل بر روی همدیگر نیز تاثیرگذار هستند سروکار داریم. اما در عمل لحاظ کردن همه این عوامل به عنوان متغیرهای مستقل در مدل، یا امکان‌پذیر نیست و یا تأثیر نامطلوب بر روی مدل خواهد داشت. پیدا کردن مدل مناسب در حیطه‌ی مربوط به انتخاب متغیر قرار می‌گیرد.
در نظر می‌گیریم مسئله انتخاب متغیر و برآورد را زمانی که مجموعه داده‌ها شامل nتا مشاهده و pتا پیش‌بینی‌کننده با بردار پاسخ $\mathbf{y} = (y_1, y_2, ..., y_n)^T \in R^n$ و ماتریس طرح $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_P) \in R^{n\times p}$ و به دنبال مدل خطی

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon. \qquad (1)$$

که در آن $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T \sim N(0, \sigma^2 I_n)$ و $\beta = (\beta_1, ..., \beta_p)^T$ هستیم. و همچنین فرض کنید که پیش‌بینی‌ها استاندار شده و پاسخ مرکزی شده باشند به طوری که

$$\sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} x_{ij} = 0, \frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1,$$

برای $j = 1, ..., p$ است. هنگامی p بزرگ است ما فرض می‌کنیم در مدل (۱) بسیاری از ضرایب رگرسیون دقیقاً صفر هستند. بدون این‌که از کلیات مسئله چیزی کم شود فرض می‌کنیم که q عنصر اولیه از بردار β غیرصفر هستند. در نظر می‌گیریم $\beta_{(1)} = (\beta_1, ..., \beta_q)^T$ و $\beta_{(2)} = (\beta_{q+1}, ..., \beta_p)^T$ در نتیجه عناصر $\beta_{(1)} \neq 0$ و $\beta_{(2)} = 0$ است. و اکنون $X_{(1)}$ و $X_{(2)}$ به ترتیب q ستون اولیه و $p-q$ باقی مانده از ستون‌های X هستند و $C = \frac{1}{n}X^T X$، که بصورت زیر بیان می‌شود.

$$\mathbf{C} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

مجموع توان دوم باقیمانده‌ها در مدل رگرسیون خطی ساده را می‌توان به صورت

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2.$$

نشان داد. هدف ما در روش حداقل‌مربعات این است که مقادیری از $\beta_0, \beta_1, ..., \beta_p$ را بیابیم که کمیت RSS را مینیمم کند.
در مواردی که متغیرهای پیش‌بین رابطه‌ی هم‌خطی داشته باشند، برآوردهای کمترین مربعات (به علت اینکه $X^T X$ بزرگ شده و در نتیجه به دلیل کوچک شدن دترمینان آن، معکوس ماتریس معنادار نخواهد شد) ضعیف عمل می‌کند.

برای رسیدن به پیش‌بینی بهتر در مواجه با هم‌خطی، از روش‌های انقباضی[1] استفاده می کنیم. در بخش بعدی

[1] Shrinkage

برخی از این روش‌ها و ویژگی‌های آنها را مرور خواهیم کرد. در بخش سوم با ذکر دو مثال ضرایب رگرسیونی را با استفاده از روش‌های ذکر شده در بخش دوم برآورد می‌کنیم. در بخش چهارم نتایج بدست آمده را تحلیل می‌کنیم و به مقایسه این روش‌ها می‌پردازیم.

۲ روش‌های انقباضی

روش‌های انقباضی از جمله روش‌های جدید در امر برآوردیابی و انتخاب متغیر در مدل‌های رگرسیونی هستند. فلسفه این روش‌ها در نظر گرفتن مقداری اریبی برای برآوردها در تلاش برای کاهش واریانس است بطوری که در نهایت میانگین مربع خطا کاهش یابد.

۱.۲ رگرسیون ستیغی

رگرسیون ستیغی[۲][۲] مانند حداقل‌مربعات، برآورد ضرایبی را جستجو می‌کند که در آن RSS کمترین مقدار شود با این شرط که این مینیمم‌سازی با جریمه‌ی انقباض $\lambda \sum_{j=1}^{p} \beta_j^2$ همراه می‌باشد.

در این روش، برآورد ضرایب $\hat{\beta}^{ridge}$، مقادیری هستند که با مینیمم کردن عبارت

$$\mathbf{Q} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2 = RSS + \lambda \sum_{j=1}^{p}\beta_j^2 \quad (۲)$$

به دست می‌آیند.

بر خلاف روش حداقل‌مربعات که فقط یک مجموعه از برآورد ضرایب را ایجاد می‌کند، رگرسیون ستیغی به ازای هر مقدار λ، مجموعه متفاوتی از برآورد ضرایب $\hat{\beta}^{ridge}$ را می‌دهد. بنابراین انتخاب مقدار λ مناسب، ضروری است.

یکی از ویژگی‌های رگرسیون ستیغی این است که جریمه‌ی $\lambda \sum_{j=1}^{p} \beta_j^2$، ضرایب را به صفر کاهش می‌دهد ولی هیچ کدام را صفر نمی‌کند مگر اینکه λ بسیار بزرگ باشد. این ویژگی ممکن است برای دقت پیش‌بینی مشکلی ایجاد نکند ولی در تفسیر مدلی که تعداد متغیرهای آن زیاد است، چالش ایجاد می‌کند.

۲.۲ روش لاسو

روش لاسو[۳][۳] به عنوان جایگزینی برای روش حداقل مربعات و با دو هدف عمده ارائه شد: هدف اول، بهبود دقت پیش‌بینی و هدف دیگر بهبود تفسیر مدل با تعیین زیر مجموعه کوچک‌تری از متغیرهای کمکی با بیشترین اثر می‌باشد. در این روش برای برآورد β ، کمیت

[۲] Ridge Regression

[۳] Lasso Regrssion

$$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{p}|\beta_j| = RSS+\lambda\sum_{j=1}^{p},|\beta_j| \quad (3)$$

نسبت به β مینیمم می‌شود. تفاوتی که این رابطه با روش رگرسیون ستیغی دارد در این است که به جای β_j^2 از $|\beta_j|$ استفاده می‌شود. در واقع از جریمه‌ی نرم ℓ_1 به جای ℓ_2 استفاده شده است. در لاسو، برخی از ضرایب به طور دقیق برابر با صفر می‌شود و این در حالتی اتفاق می‌افتد که پارامتر تنظیم کننده[۴] به اندازه کافی بزرگ باشد. در واقع می‌توان گفت که در لاسو انتخاب متغیر نیز صورت می‌گیرد، بنابراین مدل‌هایی که با روش لاسو بدست می‌آیند را بهتر می‌توان تفسیر کرد.

۳.۲ روش الاستیک نت

در مسائلی که $p > n$ انتخاب متغیر لاسو حداکثر n متغیر است. پس تعداد متغیرهای انتخاب شده به تعداد نمونه محدود می‌شود و همچنین لاسو نمی‌تواند به انجام انتخاب متغیرگروه‌بندی بپردازد این منجر به انتخاب یکی و چشم پوشی از بقیه می‌گردد. روش جایگزینی که برای حل این مسائل می توان به کار برد، روش رگرسیون الاستیک‌نت[۵][۴] می‌باشد.

برای مقادیر ثابت و نامنفی λ_1 و λ_2، الاستیک‌نت خام به شکل زیر تعریف می‌شود:

$$\mathbf{Q}(\lambda_1,\lambda_2,\beta) = |\mathbf{y}-\mathbf{X}\beta|^2+\lambda_2|\beta|^2+\lambda_1|\beta|_1 \quad (4)$$

که در آن

$$|\beta|^2=\sum_{j=1}^{p}\beta_j^2 \qquad |\beta|_1=\sum_{j=1}^{p}|\beta_j|.$$

می‌باشد و برآورد $\hat{\beta}$ در الاستیک‌نت خام با مینیمم کردن معادله (۴) بدست می‌آید.

$$\hat{\beta} = argmin_\beta \mathbf{Q}(\lambda_1,\lambda_2,\beta). \quad (5)$$

این روش را به عنوان یک روش حداقل مربعات جریمه‌دار می توان مشاهده کرد.

با در نظر گرفتن $\alpha = \lambda_2/(\lambda_1+\lambda_2)$ می‌توان برآورد الاستیک نت خام را به فرم زیر در نظر گرفت:

$$\hat{\beta} = argmin_\beta|\mathbf{y}-\mathbf{X}\beta|^2 \quad s.t \qquad (1-\alpha)|\beta|_1+\alpha|\beta|^2 \le t$$

و تابع $(1-\alpha)|\beta|_1+\alpha|\beta|^2$ را به عنوان جریمه الاستیک‌نت در نظر می‌گیریم، که ترکیب اکیداً محدب از جریمه لاسو و ستیغی است.

استفاده از این فرم از تابع جریمه علاوه بر قابلیت صفر برآورد کردن برخی از ضرایب، توانایی برابر برآورد کردن ضرایب متغیرهای که اثر یکسان روی متغیر پاسخ دارند و یا به شدت همبسته هستند را نیز دارد. به این ترتیب از این روش می‌توان برای گروه‌بندی پیش‌بین‌ها به خصوص زمانی که تعداد آنها زیاد باشد استفاده کرد.

[۴] Tuning Parameter

[۵] Elastic Net Regression

برای آنکه دقت پیش‌بینی بالاتری داشته باشیم برآورد الاستیک‌نت خام را بصورت

$$\hat{\beta}(elasticnet) = (۱ + \lambda_۲)\hat{\beta}(naiveelasticnet).$$

مقیاس‌گذاری می‌کنیم.

۴.۲ شبکه مقید

نمودارها و شبکه‌ها روش‌های معمول به تصویر کشیدن اطلاعات هستند. در زیست‌شناسی بسیاری از فرآیندهای زیستی مختلف مانند مسیرهای متابولیکی و شبکه‌های تنظیمی توسط نمودارها نشان داده می‌شوند.
در پژوهش‌های ژنومی علاقه خاصی به مسیرهای تنظیم ژن است. که روابط تنظیمی بین ژن‌ها یا محصول ژن را فراهم می‌کند. این مسیرها اغلب به هم پیوسته و به شکل شبکه هستند، که می‌تواند نشان‌دهنده یک گراف باشند. که در آن رئوس گراف ژن یا محصول ژن هستند و یال‌های گراف نشان‌دهنده برخی از روابط تنظیمی بین ژن‌هاست.

به منظور بیان این واقعیت که p متغیر توضیحی اندازه‌گیری شده در یک نمودار هستند اول ماتریس لاپلاس که مرتبط به نمودار است را معرفی می‌کنیم .
در گراف وزن‌دار $G = (V, E, W)$ یال‌ها را مطابق با $w(v_i, v_j) > ۰$ در نظر می‌گیریم ویالی وجود ندارد اگر $w(v_i, v_j) = ۰$ باشد.

- A_G، ماتریس مجاورت گراف G است که یک ماتریس $n \times n$ و متقارن است و داریم:

$$(A_G)_{i,j} = w(v_i, v_j).$$

- D_G، ماتریس قطری از درجه رئوس G می‌باشد و داریم:

$$D_G = diag(d(u_i), ..., d(u_n)), (D_G)_{i,i} = d_i.$$

- L_G، ماتریس لاپلاس است که بصورت زیر تعریف می‌شود:

$$L_G = D_G - A_G.$$

- $\mathbf{L}$، ماتریس لاپلاس نرمال شده است که ماتریسی $n \times n$ و متقارن است و بصورت زیر است:

(۶)

$$L(u,v) \begin{cases} ۱ - w(u,u)/d_u & u = v, d_u \neq ۰ \\ -w(u,v)/\sqrt{d_u d_v} & u, v \ adjacentare \\ ۰ & otherwise \end{cases}$$

به منظور دستیابی به انتخاب متغیر به‌صورت خودکار و برای محاسبه ساختار شبکه، جریمه‌ی که ناشی از ماتریس لاپلاس است در نظر می‌گیریم، این موجب همواری ضرایب شده و همچنین می‌تواند منجر به تفسیر بهتر در شناسایی ژن و زیرشبکه‌های که مربوط به متغیر پاسخ در زمینه زیست‌شناسی است، باشد.
برای هر ثابت نامنفی $\lambda_۱$ و $\lambda_۲$ معیار تنظیم شبکه مقید[۶] [۵] را تعریف می‌کنیم

[۶] Network-constrained

$$\mathbf{Q}(\lambda_1, \lambda_2, \beta) = (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)+\lambda_1|\beta|_1+\lambda_2\beta^T\mathbf{L}\beta, \quad (۷)$$

که در آن $|\beta|_1 = \sum_{j=1}^{p} |\beta_j|$ که نرم ℓ_1 است، که راه حلی برای پراکندگی و جمله دوم $\beta^T\mathbf{L}\beta$ شامل یک راه حل همواری از β در شبکه است.
برآوردگر تنظیم شبکه مقید از مینیمم کردن معادله (۷) بدست می‌آید.

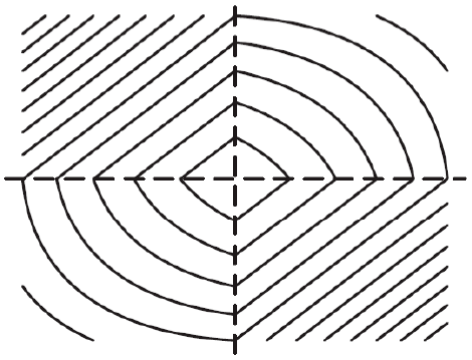
$$\hat{\beta} = argmin_\beta \mathbf{Q}(\lambda_1, \lambda_2, \beta). \quad (۸)$$

با در نظر گرفتن $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$، می‌توان برآورد شبکه مقید را به صورت زیر در نظر گرفت:

$$\hat{\beta} = argmin_\beta |\mathbf{y} - \mathbf{X}\beta|^2,$$

$$(1-\alpha)\sum_{j=1}^{p} |\beta_j| + \alpha \sum_{u\sim v} (\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}})^2 w(u,v) \leq t.$$

و تابع فوق را جریمه شبکه مقید می‌نامیم.
تابع جریمه به کار رفته در این روش از لحاظ هندسی به صورت زیر است.



شکل ۱: ناحیه‌ی جریمه در روش شبکه‌ی مقید در حالت $p = 2$

این روش همانند روش الاستیک‌نت، توانایی برابر برآورد کردن متغیرهای پیش‌بین که به شدت همبسته و یا آنهایی که اثر یکسانی روی متغیر پاسخ دارند را دارد. یعنی علاوه بر برآوردیابی و انتخاب متغیر، قابلیت گروه‌بندی آنها نیز در این روش وجود دارد.

۳ مثال های کاربردی

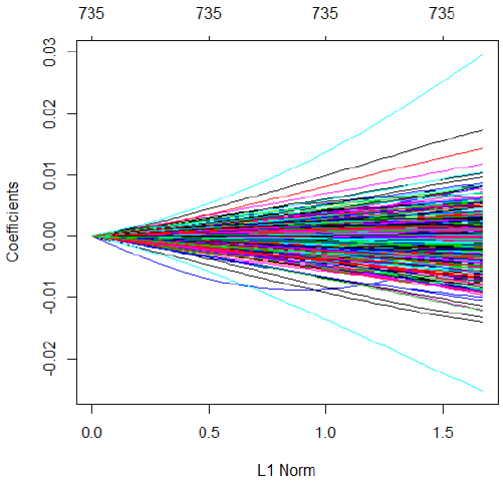
در این بخش با استفاده از دو مجموعه داده‌ی واقعی به ارزیابی عملکرد و مقایسه روش‌های پیش گفته می‌پردازیم. از آنجایی که روش‌های انقباضی در مواردی که تعداد پیش‌بین‌ها زیاد و یا حتی بیشتر از مشاهدات است کاربرد دارد و همچنین بخصوص زمانی که همبستگی بین متغیرهای مستقل زیاد باشد، عملکرد قابل قبولی از خود نشان می‌دهد، سعی بر این داریم که از داده‌های استفاده کنیم که این خصوصیات را داشته باشند.

مثال ۱ *ژن $MCM۷$ یکی از ژن های موثر در سرطان پروستات است. داده‌های بیان ژن $MCM۷$ توسط ریزآرایه‌های RNA [۷] بدست آورده شده است.*
مجموعه داده مربوط به ژن $MCM۷$ شامل ۹۰ پروب از ژن $MCM۷$ با ۷۳۵ متغیر اندازه‌گیری شده‌اند.
در جدول زیر میانگین مربعات خطا را برای داده‌های بیان ژن $MCM۷$ بدست آورده‌ایم. روش ستیغی میانگین مربعات خطای کمتری نسبت به روشهای دیگر دارد و روش الاستیک‌نت میانگین مربعات خطای کمتری نسبت به لاسو دارد.

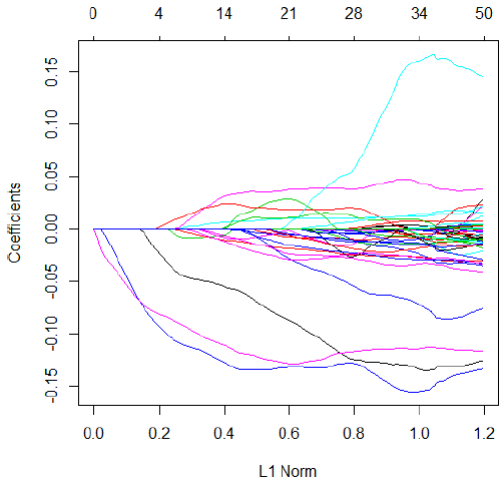
جدول ۱: نتایج حاصل از بیان ژن $MCM۷$

روش	میانگین مربعات خطای آزمون
کمترین مربعات	۰/۰۷۳
ستیغی	۰/۰۹۸۱
لاسو	۰/۰۱۶۴
الاستیک‌نت	۰/۰۱۵۹

در شکل‌های زیر برآورد ضرایب رگرسیون ستیغی، لاسو و الاستیک‌نت را برای مجموعه داده‌های بیان ژن $MCM۷$ نشان داده‌ایم. در این نمودارها، هر منحنی منطبق با برآورد ضرایب رگرسیون‌ها برای هر یک از متغیرها و به صورت تابعی از λ می‌باشد. در قسمت چپ نمودار، λ به طور ضروری صفر است و برآورد ضرایب همان برآورد عادی کمترین مربعات را نتیجه می‌دهد. با افزایش λ برآورد ضرایب به سمت صفر کاهش می‌یابد.

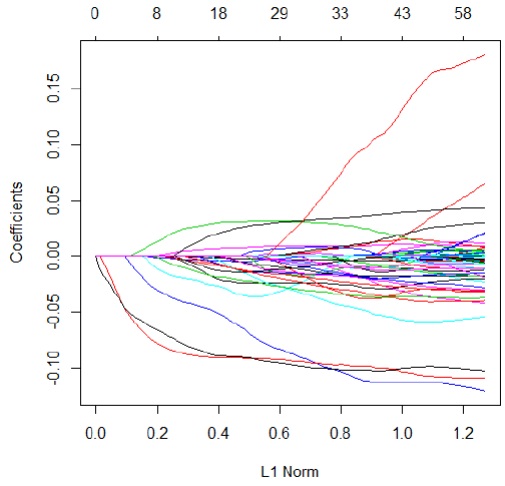


شکل ۲: برآورد ضرایب رگرسیون ستیغی



شکل ۳: برآورد ضرایب رگرسیون لاسو

با توجه به شکل‌ها مشاهده می‌کنیم که رگرسیون ستیغی ضرایب را به صفر نزدیک می‌کند ولی هیچ کدام را صفر برآورد نمی‌کند، اما رگرسیون لاسو و الاستیک‌نت با توجه



شکل ۴: برآورد ضرایب رگرسیون الاستیک‌نت

به تابع جریمه‌ی که برایشان در نظر گرفته شده، ضرایب متغیرهای که تاثیر ناچیز در مدل دارند را به طور دقیق صفر برآورد کنند و در واقع انتخاب متغیر انجام می‌دهند، بنابراین مدل بهتر تفسیر می‌شود.

مثال ۲ *آنالیز مطالعات ریز آرایه‌ای بیان ژن گلیوبلاستوما [۸].*

گلیوبلاستوما رایج‌ترین تومور بدخیم مغزی در میان بزرگسالان و همچنین کشنده‌ترین تومور در بین سایر تومورها می‌باشد. بیماران مبتلا به این تومور از زمانی که پزشکان این بیماری را تشخیص می‌دهند علی‌رغم عمل جراحی و پرتودرمانی و شیمی‌درمانی‌هایی که انجام می‌شود حداکثر تا ۱۵ ماه بیشتر زنده نمی‌مانند. داده‌های بیان ژن کلی از دو مجموعه مستقل از نمونه‌های تومورهای بالینی از $n = ۶۵$ و $n = ۵۵$ می‌باشد که از آرایه‌های $Affymetrix$ به دست آورده شده‌اند.

بر اساس آخرین پیگیری‌ها در مجموعه اول ۵۵ بیمار تنها ۵ نفر و از مجموعه دوم تنها ۴ نفر زنده مانده‌اند. مجموعه اول را به عنوان داده‌های آموزشی در نظر می‌گیریم، با استفاده از اطلاعات زمان مرگ ۵۰ بیمار مدل پیش‌بینی را ایجاد کرده‌ایم و مجموعه دوم را به عنوان داده‌های آزمون در نظر می‌گیریم، با استفاده از ۶۱ بیمار و اطلاعات زمان مرگشان عملکرد مدل پیش‌بینی را آزمون می‌کنیم.

برای اجرای آنالیز شبکه مبنی بر داده، ما داده‌های بیان ژن را با مسیر تنظیمی $KEGG(http://www.kegg.jp)$ ۳۳ ترکیب کرده‌ایم و ۱۵۳۳ ژن را در تراشه $Hu۱۳۳A$ شناسایی کرده‌ایم و هدفمان شناسایی ژن‌ها و زیرشبکه‌های مسیر $KEGG$ ۳۳ هست که با مدت زمان بقای سرطان مغز ارتباط دارد.

جدول ۲ نتایجی از روندهای متفاوت مربوط به خطای پیش‌بینی در مجموعه داده‌ها آزمون و تعداد ژن‌هایی که توسط این روندها در مجموعه آموزش انتخاب می‌شوند را نشان می‌دهد. هر دو روش الاستیک‌نت و شبکه مقید، در نتیجه مشابه‌ای باعث می‌شود که خطای پیش‌بینی نسبت به لاسو کوچکتر شود و با این حال روند شبکه مقید نسبت به لاسو و الاستیک‌نت ژن‌های بیشتری انتخاب کند و حدود نیمی از ژن‌ها (۴۴ژن) به مسیر $KEGG$ متصل شوند . لاسو در مقایسه با الاستیک‌نت که تنها یک جفت از ژن‌های متصل شده $(PRKCG \sim ITGB۷)$ را شناسایی می‌کند می‌تواند سه جفت از ژن‌های متصل شده $(FOXO۱A \sim PRKCG, PCK۱ \sim PTEN, ITGB۷ \sim SYNJ۲)$ را

جدول ۲: نتایج حاصل از آنالیز مجموعه داده‌های گلیوبلاستوما که میانگین مربعات خطای آزمون بر اساس مجموعه مستقلی از ۶۱ بیمار گلیوبلاستوما محاسبه شده است

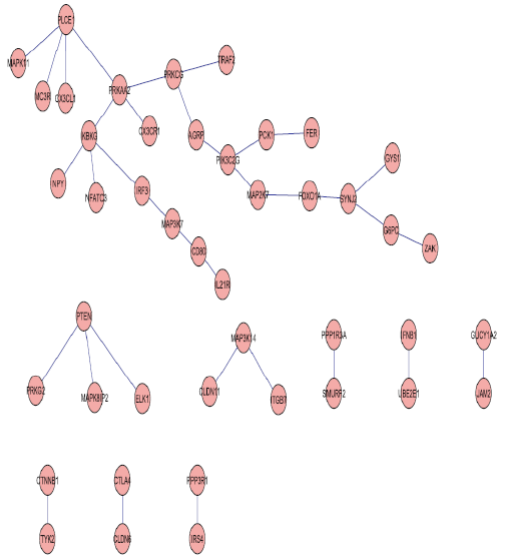
روش	میانگین مربعات خطای آزمون	تعداد ژن‌های انتخاب شده
لاسو	۱٫۱۸	۲۳
الاستیک‌نت	۱٫۰۲	۵
شبکه مقید	۱٫۰۶	۹۵

شناسایی کند .

این ژن‌ها اطلاعات زیادی درمورد مسیرها و زیر شبکه‌ها ارائه نمی‌دهند و شاید با بقای گلیوبلاستوما ارتباط داشته باشد، اما ژن‌هایی که توسط شبکه مقید شناسایی می‌شود شامل تمام ژن‌های شناسایی شده توسط الاستیک‌نت و لاسو می‌باشد.

نتایج برآمده از آنالیز شبکه مقید در واقع حاکی از آن است که امکان دارد مسیرهای متفاوتی با زمان مرگ توسط گلیوبلاستوما ارتباط داشته باشد.

شکل ۵، زیر شبکه‌های متصل شده $KEGG$ را نشان می‌دهد که توسط روند شبکه مقید پیشنهادی شناسایی شده است.



شکل ۵: زیرشبکه‌های که توسط روش شبکه مقید شناسایی شده‌اند که بر اساس نمونه‌ای از ۵۰ بیمار که ممکن است با زمان بقای بیماران گلیوبلاستوما ارتباط داشته باشد

۴ بحث و نتیجه‌گیری

با توجه به اهمیت روزافزون مبحث انتخاب متغیر در علوم مختلف مانند زیست شناسی، پزشکی و ژنتیک و..... همواره روش‌های جدیدی برای این منظور ارائه می‌گردد. در این بین استفاده از روش‌های انقباضی، بسیار مورد توجه قرار گرفته است. در این مقاله به بررسی برخی از روش‌های انقباضی به طور مختصر پرداخته‌ایم.

انتخاب متغیرهای مؤثرتر در مدل معمولاً باعث بالا رفتن دقت پیش‌بینی می‌شود. مدل‌هایی که در آنها برآورد کمترین مربعات برای همه ضرایب متغیرها محاسبه شده و عملاً انتخاب متغیری در آنها صورت نگرفته است دارای کمترین میزان دقت پیش‌بینی هستند.

با توجه به نتایج بدست آمده مشاهده می‌کنیم که رگرسیون ستیغی دارای میانگین مربعات خطای کمتری نسبت به روش کمترین مربعات دارد و همچنین با توجه به شکل۴ مشاهده می‌شود که عیب اصلی این روش غیر صفر برآورد کردن ضرایب و در نتیجه عدم انتخاب متغیر است.

بررسی نتایج نشان می‌دهد روش لاسو در مقایسه با روش ستیغی دارای دقت بیشتری است. یکی از ایرادات این روش در انتخاب متغیرهای موثر، زمانی که داده‌ها شامل گروه هایی از متغیرهای پیش‌بین به شدت همبسته‌اند، می‌باشد. روش الاستیک‌نت در مقایسه با دو روش قبلی دارای میانگین مربعات خطای کمتری است و همچنین این روش از نظر انتخاب متغیرهای موثر همبسته بهتر از روش‌های قبلی است.

یکی از راه‌های پیش‌بینی بقای بیماران مبتلا به سرطان استفاده از داده‌های بیان ژن و انتخاب ژن‌های موثر بر بقای بیماران است. مطالعه داده‌های گلیوبلاستوما نشان داد که استفاده از شبکه مقید به خوبی ژن‌های موثر را شناسایی می‌کند. بر اساس نتایج بدست آمده از مطالعه مجموعه داده گلیوبلاستوما، روش شبکه مقید در مقایسه با روش‌های لاسو و الاستیک نت از برتری قابل توجه‌ای برخوردار است (شکل ۵ را ببینید).

به طور خلاصه، این نتایج نشان می‌دهد که با در نظر گرفتن مسیرهای KEGG ، روش پیشنهادی (شبکه مقید) می‌تواند زیرشبکه‌هایی را شناسایی کند که رابطه جدی با زمان مرگ بواسطه گلیوبلاستوما دارد . برخی از این زیرشبکه‌ها توسط نتایج تصویب شده قبلی به خوبی حمایت شده است . در مقابل، ژن‌هایی که توسط الاستیک‌نت و لاسو شناسایی شده‌اند نمی‌توانند مسیرهای که ممکن ارتباط مستقیمی با خطر مرگ توسط گلیوبلاستوما دارد را نشان دهند .

با این حال ، روش شبکه مقید می‌تواند به سایر شبکه‌ها و مسیرها نیز اعمال شود. یک سوال مهم این می‌باشد که چه مسیرهایی می‌بایست در تجزیه و تحلیل داده‌های بیان ژن استفاده شود؟

این مسئله تاحدی به سوال‌های علمی که به آن پرداخته می‌شود بستگی دارد. اگر یک محقق به مسیر خاصی علاقه داشته باشد بنابراین می‌تواند روش شبکه مقید را به این مسیر خاص اعمال کند و اگر یک محقق علاقه زیادی به کاوش در مورد داده‌های خود و مسیرهای موجود داشته باشد بنابراین مجموعه بزرگی از مسیرها

مانند مسیرهای جمع آوری شده با مسیر عام : (*http* //*www.Pathwaycommons.org/pc/*) را به کار می‌برد و یا با استفاده از برخی ابزارهای موجود ساخت شبکه، شبکه‌ای از مسیرها را ایجاد می‌کند. باید به این نکته توجه داشت که روش شبکه مقید می‌تواند تمامی ژن‌های پروب شده در زیرآرایه‌ها را در برگیرد و به سادگی آنها را به عنوان گره به گراف اضافه می‌کند .

مراجع

[1] Schimek, M.G. (2003). Smoothing and Regression,approaches,computation and application.

[2] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12,55-67.

[3] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society B, 58, 267-288.

[4] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B, 67(2), 301-320.

[5] Li,C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics, 24, 1175–1182.

[6] Li, H. and Li, C. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. The Annals of Applied Statistics, 4, 1498-1516.

[7] Tye, B. K. (1999). MCM proteins in DNA replication. Annual Review of Biochemistry, 68(1), 649–686.

[8] Horvath,S. et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a novel molecular target. Proceedings of the National Academy of United States of America, 103, 17402–17407.

[9] Chiang, D. T. and Niu, S. C. (1981). Reliability of a consecutive k-out-of-n: F system. IEEE Transactions on Reliability, R-30, 87–89.