یک شبیه سازی برای تجزیه و تحلیل روش انتخاب ویژگی با استفاده از آنتولوژی ژنی برای طبقه بندی بیان ژن

چکیده

طبقه بندی مشخصات بیان ژنی یک دامنه ی تحقیقاتی اساسی است که میتواند برای روش های نوین انتقال دارو مورد استفاده قرار گیرد. یک چالش اصلی در رابطه با طبقه بندی داده های بیان ژنی تعداد کمک نمونه ها مرتبط با تعداد زیاد ژن ها است. برای حل این مشکل ، محققان الگوریتم های مختلف گزینش مشخصه را برای کاهش تعداد ژن ها پیشنهاد کرده اند. مطالعات اخیر با استفاده از شباهت معنایی بین ژن ها در آنتولوژی ژن ها (GO) به عنوان روشی برای بهبود گزینش مشخصه ، آزمایش هایی را انجام داده اند. در حالی که تعداد کمی از مطالعات وجود دارد که به چگونگی استفاده از GO ها برای گزینش مشخصه میپردازد ، هیچ مطالعه ی شبیه سازی وجود ندارد که مشخص کند که چه زمانی میتواند از گزینش مشخصه ی GO استفاده کرد. برای بررسی این موضوع ، ما یک شبیه سازی جدید را توسعه دادیم ، که مجموعه داده های طبقه های باینری را ایجاد میکند که ژن هایی که به طور متفاوت بین دو طبقه بیان شده است در GO دارای رابطه هایی زیر لایه ای هستند. این امکان به وجود می آید تا تاثیر فاکتور های مختلف مانند همبندی های نسبی ژن های اصولی در GO ، مقدار میانگین جدایی بین ژن هایی که به طور متفاوت بیان شده اند که توسط δ نشان داده میشود ، و تعداد نمونه های تمرینی را مورد بررسی قرار داد. نتایج شبیه سازی ها ما بیان دارد که همبندی در GO میان ژن هایی که به طور متفاوت بیان شده است برای شرایط زیستی فاکتور اساسی برای تعیین کارایی گزینش مشخصه ی GO است. به طور خاص ، مادامی که همبندی ژن هایی که به طور متفاوت بیان شده است افزایش پیدا میکند ، بهبود صحت طبقه بندی ها نیز افزایش پیدا میکند. برای کمی سازی این مفهوم همبندی ، ما یک

اندازه گیری به نام سطح تفسیر شرایط زیستی BCAL(G) را تعریف کردیم که G گراف ژن هایی است که به طور متفاوت بیان شده است. نتیجه ی اصلی ما با توجه به گزینش مبتنی بر GO به صورت زیر است :

(1) صحت طبقه بندی ها هنگامی که $BCAL(G) \geqslant 0.696$ افزایش پیدا میکند ؛ (2) صحت طبقه بندی ها هنگامی که $BCAL(G) \leqslant 0.389$ کاهش پیدا میکند ؛ (3) بهبود صحت به صورت حاشیه ای هنگامی که $0.389 < BCAL(G) < 0.696$ و $\delta < 1$ است اتفاق می افتد ؛ (4) مادامی که تعداد ژن ها در شرایط زیستی افزایش پیدا میکند و بیشتر از 50 میشود و $\delta \geqslant 0.7,$ ، بهبود از گزینش مبتنی بر مشخصه کاهش پیدا میکند ؛ و (5)ما استفاده از روش مبتنی بر GO را هنگامی که شرایط زیستی کمتر از 10 ژن دارد را پیشنهاد نمی کنیم. نتایج ما از مجموعه داده هایی استخراج شده است که با استفاده از RMA (میانگین چند آرایشی دقیق) پیش پردازش شده و موارد به گونه ای هستند که δ بین 0.3 و 2.5 است و سایز نمونه های تمرینی بین 20 تا 200 است ، ازین رو نتایج ما به این مشخصه های محدود هستند. به طور کلی ، این شبیه سازی ها ابتکاری بوده و به این سوال پاسخ میدهد که چه زمانی باید از حالت گزینش مشخصه SoFoCles برای طبقه بندی به جای اندازه گیری های مبتنی بر آمار ، استفاده کرد.

محتویات اطلاعاتی نسبت به GO به صورت ذاتی هستند ، ازین رو هیچ مجموعه ی خارجی از اطلاعات وجود ندارد که نیاز به محاسبه داشته باشد. با استفاده از این اطلاعات ذاتی ، یک نسل بعدی از عبارت ها میتواند به عنوان رخدادی از اجدادش در نظر گرفته شود. ما از این قرار داد استفاده کردیم که یک عبارت ولد و سرپرست خودش است ، زیرا این یک تنظیمات پارامتر پیش فرض در متلب بود ، که متلب پلتفرم نرم افزاری است که برای این بررسی مورد استفاده قرار گرفته است. میتوان اندازه گیری محتویات اطلاعاتی را گونه ای نرمال سازی کرد که فقط مقادیری بین صفر و یک گرفته که این کار توسط تقسیم محتویات اطلاعاتی یک برگ انجام میشود. به این علت که برگ ها دارای بیشترین محتویات اطلاعاتی هستند ، این موضوع محتویات اطلاعاتی نرمال شده ی آن ها را یک میکند . اقتباس ریاضیاتی این مفهوم به صورت زیر است :

$$IC(leaf) = -\log(p(leaf)) = -\log\left(\frac{1}{n_r}\right) \quad (2)$$

$$IC_{norm}(t) = \frac{IC(t)}{IC(leaf)} = \frac{\log\frac{n_t}{n_r}}{\log(\frac{1}{n_r})} = 1 - \frac{\log(n_t)}{\log(n_r)} \quad (3)$$

ایده ی محتویات اطلاعاتی منجر به روش های متعددی برای محاسبه ی مشابهت معنایی بین دو عبارت GO میشود . ایده ی مشابهت معنایی برای محاسبه ی مقدار مرور روش های مشابهت معنایی برای آنتولوژی های زیستی دارویی است. یک اندازه گیری برای مشابهت معنایی روش رسنیک است ، و این روش با تخصیص مشابهت های معنایی بین دو عبارت کار میکند تا محتویات اطلاعاتی پایین ترین جد مشترک خودشان باشند.

فرمول ریاضی شباهت معنایی رسنیک بین دو عبارت t_1 و t_2 به صورت زیر است :

$$\text{R-}sim_{norm}(t_1, t_2) = \frac{\max_{t \in S(t_1,t_2)}[IC(t)]}{IC(leaf)} \quad (4)$$

که $S(t_1, t_2)$ مجموعه ی مشترک جد ها برای t_1 و t_2 است.

آخرین پیش نیاز برای محاسبه ی شباهت معنایی بین دو ژن راهی است برای مقایسه ی چندین عبارت به صورت همزمان. این حالت مورد نیاز است به این علت که ژن های GOA میتوانند روی مجموعه ای از عبارات GO توضیح داده شود. با دو ژن ، ما میتوانیم شباهت بین دو مجموعه عبارات GO را مطابق با دو ژن به صورت یک ماتریس نشان بدهیم :

$$SIM(a, b) = \begin{bmatrix} sim_{1,1} & sim_{1,2} & \cdots & sim_{1,N_b} \\ \vdots & \vdots & \ddots & \vdots \\ sim_{N_a,1} & sim_{N_a,2} & \cdots & sim_{N_a,N_b} \end{bmatrix} \quad (5)$$

که N_a تعداد عبارات GO برای ژن a و N_b تعداد عبارات GO برای ژن b است.

شباهت بین ژن A و ژن b میتواند توسط $Sim_{MAX}(a, b)$ مشخص شود ، که بیشترین مقدار ماتریس $SIM(a, b)$ را میابد.

$$Sim_{MAX}(a, b) = \max_{i,j}(sim_{i,j}) \qquad (6)$$

در این بخش ، ما ایده ی محتویات اطلاعاتی را معرفی کردیم که روشی است برای کمی سازی مقدار اطلاعات ذاتی نسبت به یک عبارت GO . ما سپس در رابطه با روشی بحث کردیم تا بتوانیم شباهت معنایی بین دو عبارت GOرا محاسبه کنیم. این روش شباهت معنایی بین دو عبارت GO را با صورت محتویات اطلاعاتی پایین ترین جد مشترک در نظر میگیرد. این ایده میزان اطلاعات مشترک بین این دو عبارت را نشان میدهد. نهایتا ، ما یک روش برای مقایسه شباهت معنایی بین دو ژن را با یافتن بیشترین شباهت بین عبارت های GO نسبت داده شده به این ژن ها را ذکر کردیم.

3.روش ها

هدف این بخش توضیح دادن روش های مورد استفاده در شبیه سازی ما است. ما در ابتدا چگونگی اینکه شبیه سازی ها داده های بیان ژن ها را از GO ایجاد میکنند را مورد بحث قرار میدهیم. ما از ورژن 1.1807 GO استفاده کرده ایم که ما آن را در تاریخ 03/01/2011 (mm/dd/yyyy)(از سایت http://www.gene- ontology.org/GO.downloads.ontology.shtml. دانلود کرده ایم. ما GOAرا در تاریخ 03/8/2011 از سایت ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/HUMAN/gene_association. goa_human.gz دانلود کردیم. برای شبیه سازی ما ، ما از گونه های انسانی برای GOA استفاده کردیم و این نسخه از GOA دارای 18141 ژن است. برای GO ، ما تحلیل را نسبت به آنتولوژی پردازش های زیستی محدود کردیم. ما همچنین تفسیر های IEA را که تنها تفسیر های به دست آمده به صورت خودکار هستند را استثنا قرار داده ایم. ما این انتخاب ها را با توجه به GO انجام داده ایم

تا با مرجع 38 مطابقت داشته باشد. مجموعه داده های بنیادی برای تولید داده های بیان ژن ها GDS2771 است که از مرجع بیان ژن ها (GEO) به دست آمده است. این مجموعه داده توسط Spira و همکارانش به دست آمده است و از سلول های پوششی مسیر هوایی سیگاری های با سرطان و سیگاری های بدون سرطان با استفاده از میکروآرایه های HG-U133A. به دست آمده است. او و همکارانش از الگوریتم میانگین دقیق ریز آرایه ها برای به دست آوردن داده های سطح پروبی استفاده کردند. 192 نمونه در این مجموعه داده ها وجود دارد ؛ 90 نمونه بدون سرطان و 102 نمونه با سرطان هستند. دلیل اولیه برای انتخاب این مجموعه داده سایز آن است. بعد از این که ما روند تولید داده ها را مورد بحث قرار دادیم ، ما یک نسخه ی کمی اصلاح شده از SoFoCles, را معرفی میکنیم که تکنیک گزینش مشخصه ای است که از GO برای گزینش ژن ها استفاده میکند. بعد از این موضوعات مهم ، ما روش شناسی کلی آزمایشی خود را توضیح میدهیم. ما این بخش را با اندازه گیری هایی که ایده ی همبندی برای شرایط زیستی و یا مشکلات زیستی را کمی سازی میکند و توسط روش ما ایجاد شده است ، خاتمه میدهیم.

3.1 روش های شبیه سازی

ما تحلیل خودمان را به مواردی محدود میکنیم که فقط دو گروه وجود دارد ، یک گروه کنترل و یک گروه آزمایشی. ما از این عبارات به طور کاملا بی ربط استفاده میکنیم ؛ زیرا شبیه سازی ما باید روی دو طبقه از داده های مسائل بیان ژن ها اعمال شود. در این حالت ، دو نکته ی اساسی وجود دارد که باید در رابطه با آن بحث شود : (1) الگوریتم 1 که راهی برای تعریف یک گروه ژن های متمایز کننده بین طبقه ی آزمایشی و طبقه ی کنترلی از GO است ؛ و (2) یک روش برای ایجاد داده های بیان ژن ها با استفاده از این گروه از ژن ها.

الگوریتم 1

α % کمترین تعداد از ژن هایی است که باید مشخص شوند

$\%\ \beta$ سرحد محتویات اطلاعاتی است.

$\%\ G_{enes}$ لیستی از n نماد ژن هاست.

$\%\ g_i$ نماد ژن در شاخص i است

$\%$ GO لیست تمام عبارات GO است

$\%\ Annotated = \{(g, t) \mid g \in G_{enes} \wedge t \in GO \wedge t$ نسبت به g در GOA تفسیر شده است }

$\%\ \Delta$ لیست خروجی ژن هایی است که طبقه های کنترل و آزمایش را از یکدیگر تفکیک میکنند.

$$\Delta \leftarrow \emptyset$$

در حالی که $|\Delta| \leqslant \alpha$ گام های زیر را انجام میدهیم :

عدد صحیح $i \leftarrow (0, n-1)$

$$\Delta \leftarrow \Delta \cup g_i \in G_{enes}$$

$$GOIDs \leftarrow \{t \mid (g_i, t) \in Annotated \wedge IC_{norm}(t) \geqslant \beta\}$$

برای تمام $t \in GOIDs$ گام های زیر را انجام میدهیم :

$$\Delta \leftarrow \Delta \cup \{g \mid (g, t) \in Annotated\}$$

پایان حلقه ی تکرار

پایان حلقه ی شرطی

مقدار $|\Delta|$ را باز گردان

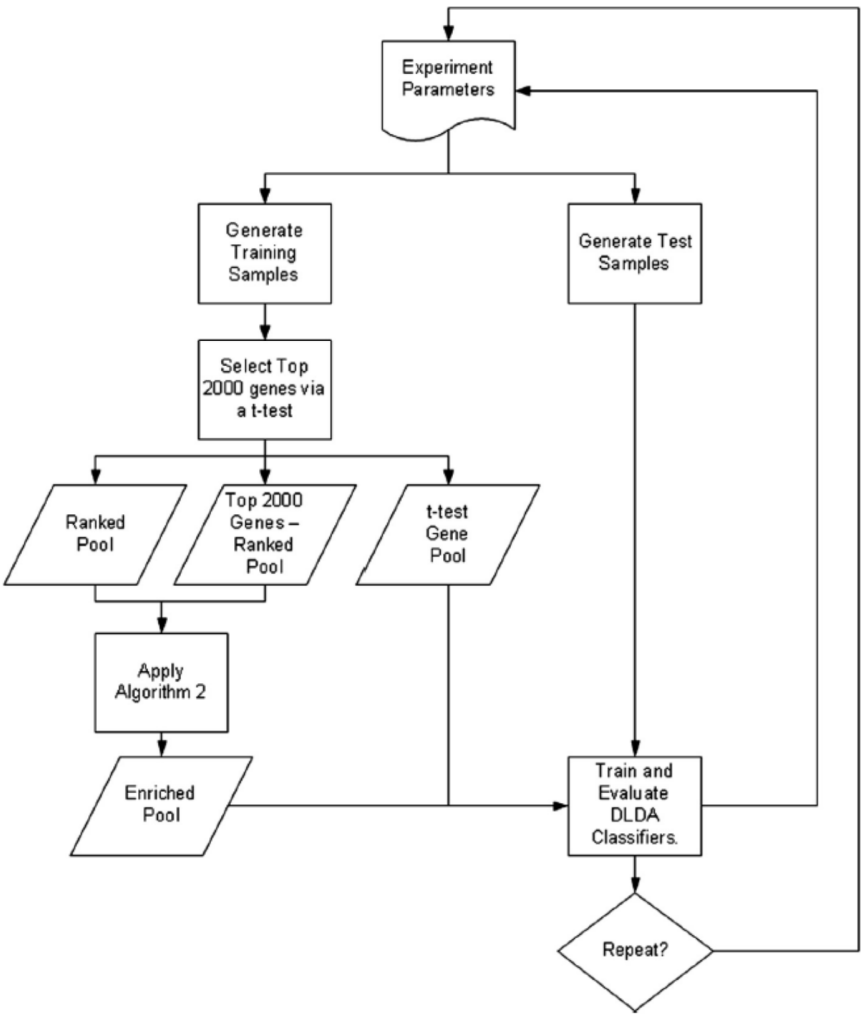
خروجی الگوریتم 1 ، Δ_1 ، مجموعه از ژن ها است که به طور متفاوتی بیت طبقه های کنترلی و آزمایشی بیان شده اند. این مجموعه ی Δ_1 گام تولید داده ها است که ما بعدا در رابطه با آن بحث خواهیم کرد. ما اکنون با خواننده های درکی د رابطه با چگونگی کارایی این الگویتم میدهیم. الگوریتم 1 با انتخاب یک ژن g_i از میان G_{ene} آغاز میشود ، که لیستی از تمام ژن ها است. این ژن سپس به Δ_0 الحاق میشود. تمام عبارت های GO که روی g_i با محتویات اطلاعاتی حد اقل β تفسیر میشود ، در مجموعه ی GOID ها قرار میگیرد. سپس ، برای هر عبارت GO به صورت t در مجموعه ی GOID ها ، ما تمام ژن های تفسیر شده روی t را میابیم ، و تمام این ژن ها را به Δ_1 الحاق میکنیم. این روند تا زمانی که سایز Δ کوچکتر و یا برابر با پارامتر α_0 است ادامه پیدا میکند. اکنون ما Δ_1 را تعریف کرده ایم ، ما میتوانیم در رابطه با تولید داده های بیان ژن ها با استفاده از این مجموعه بحث کنیم.

3.1.1 تولید داده های بیان ژن ها

روش ما برای تولید داده های بیان ژن ها بر یک شبیه سازی واقعی بیان ژن ها مبتنی است که توسط Singhal و همکارانش ارائه شده است. این روش مجموعه ای از مشخصات واقعی بیان ژن ها را به عنوان بنیادی در نظر گرفته و سپس سه سطح نویز را برای ایجاد مشخصات جدید بیان ژن ها اضافه میکند. در مطالعه ی ما ، ما یک طبقه ی کنترلی و یک طبقه آزمایشی را با هم مقایسه میکنیم. داده های بنیادی کنترلی ما از نمونه های غیر سرطانی مجموعه اطلاعاتی GEO GDS2771 است. در حالی که این مجموعه اطلاعاتی دارای دو طبقه است ، ما فقط از نمونه های غیر سرطانی استفاده میکنیم زیرا ما باید ویژگی های گروه کنترلی را کنترل کنیم تا به طور کار آمد گزینش ویژگی مبتنی بر GO را مورد مطالعه قرار دهیم. دو پارامتر ، Δ و δ_1 ، طبقه ی آزمایشی را توصیف میکند. اولین پارامتر ژن هایی را تعریف میکند که واقعا به طور متفاتی بین طبقه های کنترلی و آزمایشی بیان شده اند. پارامتر دوم جدایی میانگین در بیان ژن های واقعی بین طبقه های کنترلی و آزمایشی را برای تمام ژن های Δ نشان میدهد. به این علت که GDS2771 با استفاده از RMA تحت پیش پردازش قرار

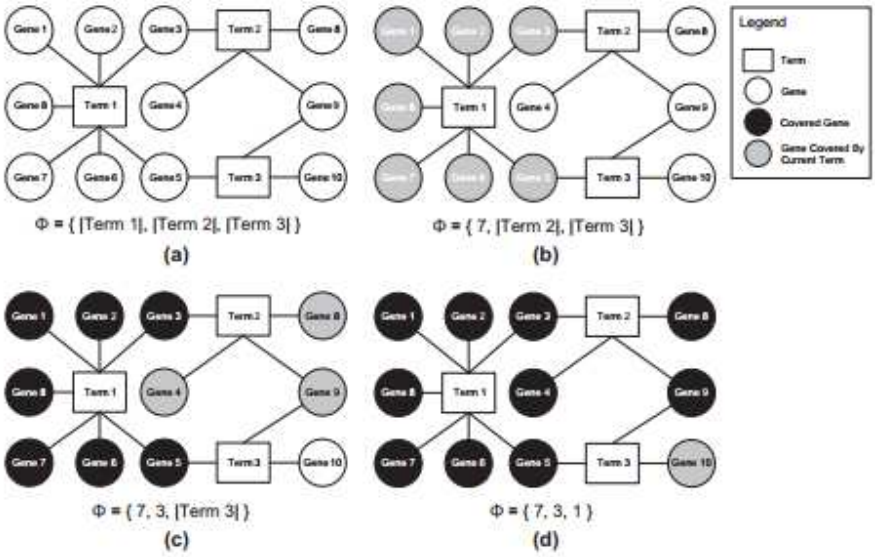
گرفته بود ، δ متناظر با واحد هایی است که در فضای بیان RMA تحت پیش پردازش قرار گرفته اند. تمام دیگر ژن ها دارای افزایش و یا کاهش رندوم در سطح بیان بین داده های بنیادی طبقه های کنترلی و آزمایشی هستند. این تغییر رندوم در بیان برای هر کدام از ژن ها توسط متغیر رندوم نرمال با میانگین صفر و یک انحراف استاندارد 0.1 تعریف میشود. ازین رو برای Δ و δ خاص ، ما یک داده ی بنیادی آزمایشی از داده های بنیادی گروه کنترلی ایجاد میکنیم.

برای تولید داده های جدید مصنوعی برای بیان ژن ها ، ما سه منبع نویز را به داده های بنیادی اضافه میکنیم. سه منبع نویز که توسط Singhal و همکارانش تعریف شده اند به صورت زیر هستند : (1) تغییر های فنی سیستمی و یا تغییر های درون آرایه ای ؛ (2) تغییر پذیری های فنی رندوم و یا تغییر پذیری های میان آرایه ای ؛ و (3) تغییر پذیری های زیستی.



شکل 1 یک مرور نسبت به اجرای آزمایشی. Δ, δ_i داده شده و تعداد نمونه های آزمایشی به ازای پارامتر های طبقه ، ما نمونه های آزمایشی و تست را ایجاد میکنیم. یک تست t دو نمونه ای برای ترتیب دهی به ژن ها استفاده شده است ، که ژن هایی با رتبه ی بالاتر از 2000 حذف میشوند. ژن هایی با بیشترین مقادیر r محض در تست t به عنوان گروه های رتبه بندی شده استفاده میشود ، و ژن هایی با بیشترین مقدار محض 2r در تست t به عنوان گروه تست t انتخاب میشود. گروه غنی شده نیز از گروه های رتبه بندی شده از الگوریتم 2 ساخته میشود. دو طبقه بندی کننده ی DLDA مورد تمرین قرار میگیرد که یکی از گروه غنی شده استفاده

میکند و یکی از گروه t . این طبقه بندی کننده ها روی نمونه های تست ارزیابی میشود. این روند 20 بار برای یک مجموعه داده شده از پارامتر ها تکرار میشود.



شکل 2

جزئیات ریاضی این روش و بحث های بیشتر از سه نوع نویز در ضمیمه ی A وجود دارد.

3.2 اجرای شبیه سازی

ما اکنون بعضی از تنظیم های پارامتر ها را توضیح داده و مورد بحث قرار میدهیم که برای شبیه سازی استفاده میشود و روش آزمایش خودمان را توضیح میدهیم. در آزمایش های ما ، ما صحت دو طبقه بندی کننده ی تحلیل متمایز کننده قطری خطی (DLDA) که روی داده های ایجاد شده از دو طبقه تمرین داده میشود را مقایسه میکنند. یک طبقه بندی کننده با استفاده از یک زیر مجموعه از ژن ها که از نظر آماری اهمیت دارند تمرین داده میشود. دومین طبقه بندی کننده نیز با استفاده از مجموعه ای از ژن ها تمرین داده میشود که از نظر آماری نیمی از اهمیت ژن های مهم را داشته و از نظر معنایی نیز نصف اهمیت را دارا هستند. طبقه بندی کننده ی اول که ما آن را به عنوان طبقه بندی کننده ی آماری یاد میکنیم ، و طبقه بندی کننده دوم که ما به

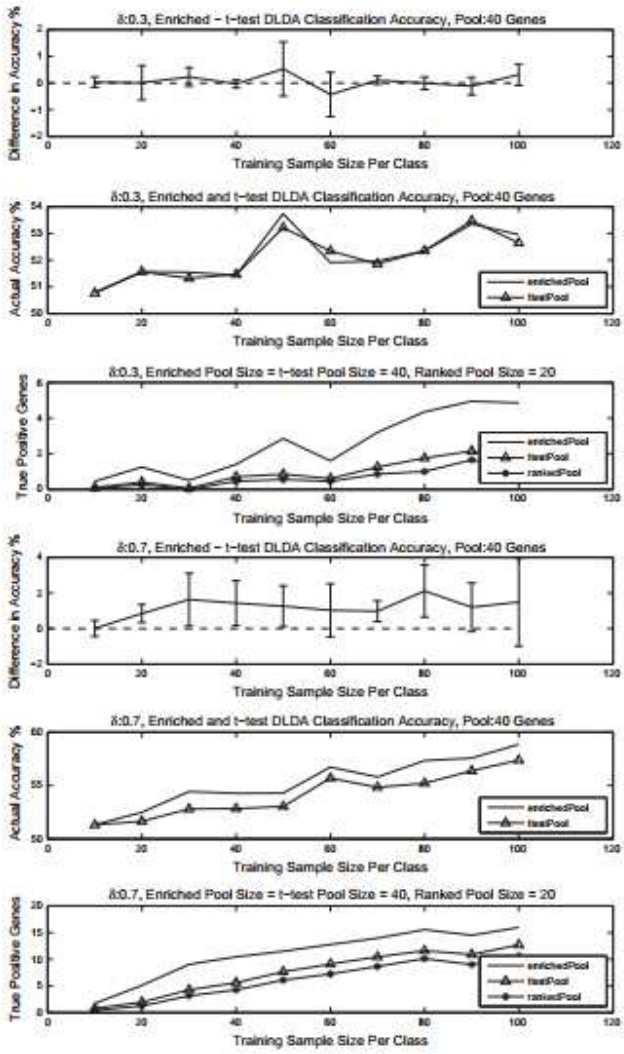
آن طبقه بندی کننده غنی شده میگوییم. تعداد ژن های استفاده شده برای تمرین هر دو طبقه بندی کننده مشابه است. با استفاده از این روش ، ما داریم تست میکنیم که آیا یک گروه از ژن ها که از نظر معنایی و آماری نیمی از اهمیت ژن های مهم را دارا هستند موثر تر از استفاده از ژن های مهم آماری به تنهایی است ، یا خیر.

سوال اصلی در این مطالعه این است : استفاده از گزینش مبتنی بر GO چه زمانی مفید است ؟ برای کمک به پاسخ به این سوال ، ما یک اندازه گیری به نام سطح تفسیر شرایط زیستی (BCAL(g)) را تعریف کردیم. ما BCAL(g) را به طور دقیق در بخش 3.3 معرفی میکنیم. اما ، در این جا فقط همین کافی است که ایده ی مبنایی BCAL(g) را درک کنید. این پارامتر از مجموعه ای از ژن های Δ محاسبه شده و مقداری بین صفر و یک را دریافت میکند. هر این مقدار به یک نزدیک تر باشد ، ژن های Δ به طور متمرکز تر در GO قرار دارند. این بیان دارد که ، هر چه این مقدار به یک نزدیک تر باشد ، بهبود بیشتری را از گزینش مشخصه ی مبتنی بر GO انتظار داریم. برای این که ببینیم $BCAL(G)$ چگونه بر گزینش مبتنی بر GO تاثیر دارد ، ما 10 حالت زیستی متفاوت را با مقادیر $BCAL(G)$ متفاوت ایجاد کرده ایم. به یاد بیاورید که Δ مجموعه ای از ژن ها را تعریف میکند که به طور متفاوتی بین گروه کنترل و آزمایش بیان شده اند. هر Δ دارای یک $BCAL(G)$ متناظر است که با آن مرتبط شده است. برای ایجاد یک Δ اولیه ، ما الگوریتم 1 را با $\alpha = 30$ و $\beta = 1$ اعمال کرده ایم. این پارامتر ها موجب به وجود آمدن $|\Delta| = 36$ میشود ، که $|\Delta|$ نشان دهنده ی تعداد ژن ها در Δ است.

ما ده حالت زیستی جدید را با اصلاح این Δ اولیه ایجاد کرده ایم ، که آن را با Δ_0 نشان میدهیم. برای اصلاح Δ_0 ما یا ژن ها را اضافه میکنیم یا آن ها را کم میکنیم. ما ژن ها را اضافه میکنیم تا مقدار متناظر $BCAL(G)$ را کم کرده و یا آن ها را کم میکنیم تا مقدار متناظر $BCAL(G)$ را اضافه کنیم. روش ما برای اضافه کردن ژن ها به Δ_0 به صورت مقابل است : (1) نماد ژن ها را به صورت لغت نامه ای ترتیب بندی میکنیم ؛ و (2) ژن ها را با پایین ترین رتبه به اضافه میکنیم تا به $BCAL(G)$ برسیم. در بسیاری از حالات ، ما باید بعضی از ژن ها را از Δ_0 حذف کنیم تا به $BCAL(G)$ هدف برسیم. ما این شرایط زیستی را به صورت Δ_1 ... Δ_{10} نام گذاری

میکنیم. این حالات زیستی دارای مقادیر $BCAL(G)$ به صورت $\{1.000, 0.892, 0.785, 0.696, 0.584, 0.488, 0.389, 0.291, 0.182, 0.086\}$ هستند. با وجود این که ایجاد کردن یک حالت زیستی با $BCAL(G) = 1.000$ دقیق کار سختی است ، اما هدف ما شروع با $BCAL(G)$ و کاهش آن در هر حالت زیستی بعدی به اندازه ی 0.10 بود. تعداد ژن ها در هر مرحله ی حالت زیستی به ترتیب $\{23, 26, 31, 35, 36, 39, 36, 38, 38, 37\}$ است. برای هر حالت زیستی $\Delta_i \text{ for } i \in \{1, 2, \ldots, 10\}$ ، ما دو پارامتر δ و تعداد نمونه های تمرینی برای هر طبقه را تغییر دادیم. مقدار های δ که ما بررسی میکنیم $\{0.3, 0.5, 0.7, 0.9\ 1.1, 1.3, 1.5, 1.7, 1.9, 2.1, 2.3, 2.5\}$ هستند. تعداد نمونه های تمرینی در هر طبقه نیز $\{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$ میباشد. به این علت که دو طبقه وجود دارد تعداد کل نمونه های تمرینی از 20 تا 200 میباشد. برای هر Δ_i ما 120 آزمایش انجام دادیم ؛ این مقدار ها ازین موضوع ایجاد شده است که 12 مقدار δ و 10 سایز متفاوت برای نمونه وجود دارد. هر آزمایش 20 بار تکرار شده است ، و ما میانگین 20 تکرار را برای به دست آوردن میانگین تفاوت در صحت طبقه بندی ها بین طبقه بندی کننده غنی شده و طبقه بندی کننده آماری ، محاسبه میکنیم. ما به طبقه بندی کننده ای که از GO استفاده میکند غنی شده میگوییم ازین رو که ژن های آماری از نظر معنایی توسط GO غنی شده اند. برای تمام آزمایش ها ، ما تعداد ژن های انتخاب شده را روی 40 ثابت کرده ایم.

یک موضوع مهم دیگر در رابطه با $BCAL(G)$ این است : چگونه تعداد ژن ها در Δ روی گزینش مشخصه مبتنی بر GO تاثیر میگذارد ؟ برای پاسخ به این سوال ، ما روی مواردی متمرکز میشویم که $BCAL(G) = 1.000$ است زیرا ما میتوانیم به طور کار آمدی تعداد ژن ها را کنترل کنیم. ما سپس 10 حالت زیستی جدید را ایجاد کردیم : $\{\Delta_{11}, \Delta_{12}, \Delta_{13}, \Delta_{14}, \Delta_{15}, \Delta_{16}, \Delta_{17}, \Delta_{18}, \Delta_{19}, \Delta_{20}\}$. تعداد ژن ها در هر کدام از این حالت های زیستی به ترتیب $\{5, 10, 50, 100, 150, 200, 250, 300, 350, 400\}$. مقدار های مشابه برای δ و نمونه های تمرینی در هر طبقه مانند بالا در این آزمایش استفاده شده است.



شکل3

شکل 1 گام های یک اجرای آزمایشی را نشان میدهد .برای یک مقدار مشخص از δ ، ما 2000 گروه کنترل و 2000 گروه آزمایشی نمونه های تمرینی ایجاد میکنیم. یک روند مشابه نیز برای ایجاد مجموعه ی تست اعمال میشود. با در نظر داشتن این که تمام پارامتر های دیگر به صورت ثابت است ، ما یک جایگشت نمونه های تمرینی را با تعداد نمونه های مشخص شده توسط سایز نمونه های تمرینی در هر طبقه ایجاد میکنیم. برای مثال ، در صورتی که سایز نمونه ی تمرینی در هر طبقه 20 باشد ، سپس یک نمونه از مشخصات 40 بیان ژن با 20 مشخصه از گروه کنترل و 20 تا از گروه آزمایشی ایجاد میشود.

برای رتبه بندی این ژن ها ما یک تست t دو نمونه ای را برای هر ژن اعمال میکنیم. برای کاهش فضای جست و جو ، فقط 2000 ژن با بیشترین مقدار محض تست t حفظ میشود ، در حالی که بقیه ژن ها فیلتر شده و حذف میشود.

بعد از این که 2000 ژن بالایی انتخاب میشود ، ما سه گروه از ژن ها را ایجاد میکنیم ، گروه رتبه دار ؛ گروه تست t و گروه غنی شده. گروه رتبه دار شامل r های بالا است که توسط مقدار محض تست t رتبه بندی شده اند. گروه تست t شامل ژن هایی با 2r بالا هستند که توسط مقدار محض تست t رتبه بندی شده است. برای تمام آزمایش های ما $r = 20$ است. گروه غنی شده نیز توسط الگوریتم 2 ایجاد شده است.

الگوریتم 2 . الگوریتم غنی سازی

$\%\ rankedPool$ ژن های r های بالا است که توسط مقدار محض تست t رتبه بندی شده

$top2000ttest - rankedPool =$ دیگر گروه ها

$\%\ similarity[i] = 0,\ \forall i \in$ دیگر گروه ها

برای تمام $g \in rankedPool$ گام های زیر را انجام بده

برای تمام $g' \in otherPool$ گام های زیر را انجام بده

$$sim_o = Sim_{MAX}(g, g')$$

در صورتی که $similarity[g'] < sim_c$:

$$similarity[g'] = sim_o$$

پایان حلقه ی شرطی

پایان حلقه ی تکرار دوم

پایان حلقه ی تکرار اول

% گروه دیگر (other pool) را توسط شباهت به صورت نزولی منظم کرده و سپس توسط مقدار محض تست t به صورت نزولی تنظیم کنید.

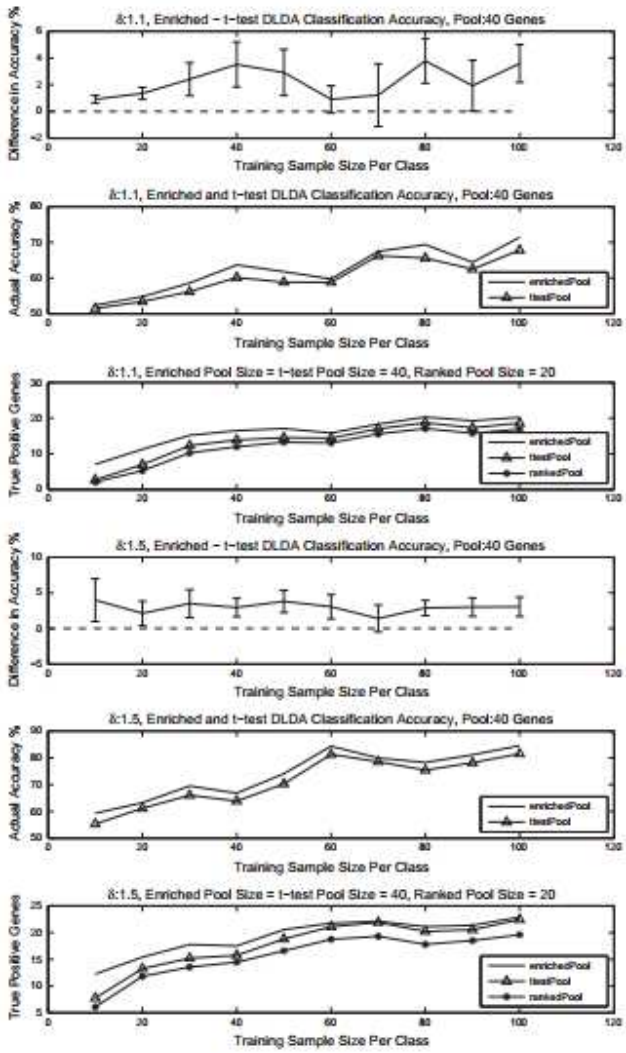
$$simPool = sort(otherPool, similarity)$$ %

گروه غنی شده = $\cup \{g \mid g\text{'s index in } simPool \leqslant r\}$ گروه رتبه دار

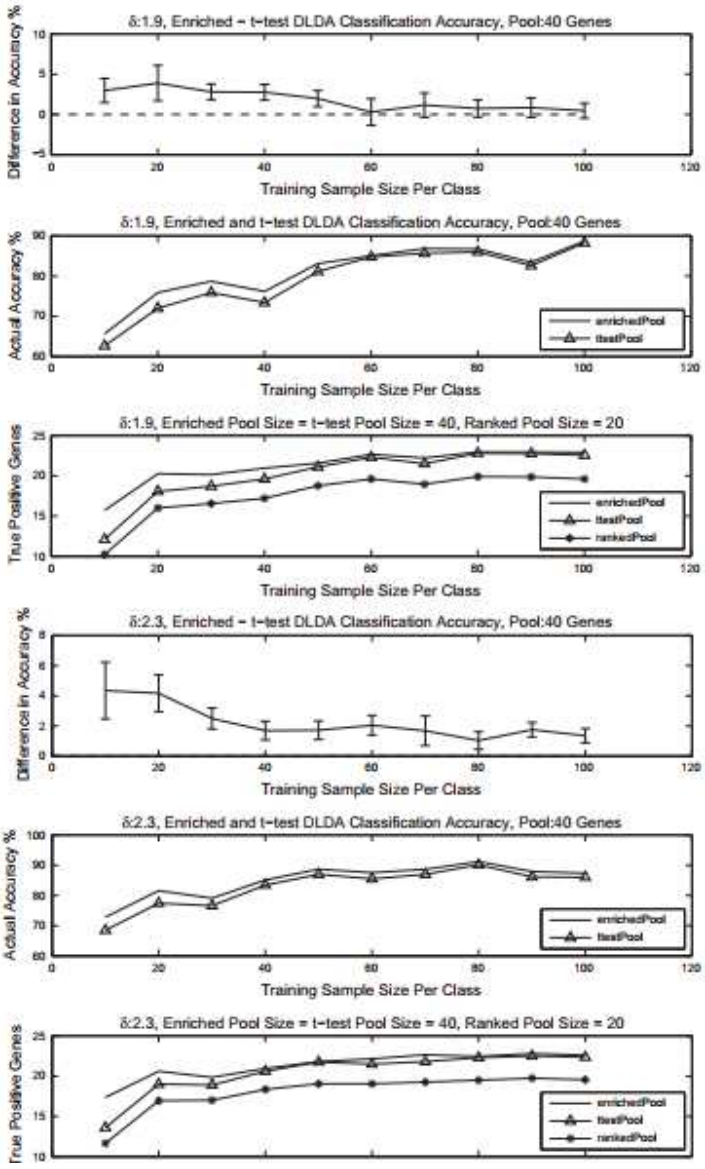
گروه غنی شده را باز گردان.

الگوریتم 2 r ژنی را که از نظر معنایی با ژن های موجود در گروه رتبه دار مشابه هستند را پیدا میکند. الگوریتم 2 بر اساس روند غنی سازی است که در مرجع 38 ارائه شده است. ما نسبت به اجرای اصلی توسط استفاده از مقدار محض t-test در قسمت منظم سازی ، متفاوت عمل میکنیم.

این موضوع در هنگام استفاده از Sim_{MAX}، مهم است زیرا بسیاری از جفت های ژن ها به صورت مشابهت معنایی برابر داده میشود. برای مثال ، هر جفت از ژن ها دارای یک تفسیر مشترک عبارت برگی هستند دارای شباهت معنایی برابر میباشند.


δ:1.1, Enriched – t–test DLDA Classification Accuracy, Pool:40 Genes
Difference in Accuracy %
Training Sample Size Per Class
δ:1.1, Enriched and t–test DLDA Classification Accuracy, Pool:40 Genes
Actual Accuracy %
enrichedPool
testPool
Training Sample Size Per Class
δ:1.1, Enriched Pool Size = t–test Pool Size = 40, Ranked Pool Size = 20
True Positive Genes
enrichedPool
testPool
rankedPool
Training Sample Size Per Class
δ:1.5, Enriched – t–test DLDA Classification Accuracy, Pool:40 Genes
Difference in Accuracy %
Training Sample Size Per Class
δ:1.5, Enriched and t–test DLDA Classification Accuracy, Pool:40 Genes
Actual Accuracy %
enrichedPool
testPool
Training Sample Size Per Class
δ:1.5, Enriched Pool Size = t–test Pool Size = 40, Ranked Pool Size = 20
True Positive Genes
enrichedPool
testPool
rankedPool
Training Sample Size Per Class

شکل 4



شکل 5

Sim_{MAX} مورد استفاده قرار گرفته است زیرا با داده های بیان ژن ها به خوبی عمل میکند. به علاوه ، در مطالعه ای توسط christoudis و همکارانش ، به طور با ثباتی عمل کرده و اجرای آن ساده است.

الگوریتم 3 . الگوریتم BCAL

%ژن پوشش داده شده توسط (t) = $\{g \mid (g, t) \in E\}$

% جدا شده = $\{ g \mid g \in \Delta \wedge g$ به هیچ $t \in Terms$ تفسیر نشده است

$$\forall i, \phi_i = 0$$

پوشش داده شده = $\emptyset$

در حالی که $|covered| < |\Delta| - |Isolated|$ گام های زیر را انجام بده

$$t \leftarrow argmax_t(|genesCoveredBy(t) - covered|)$$

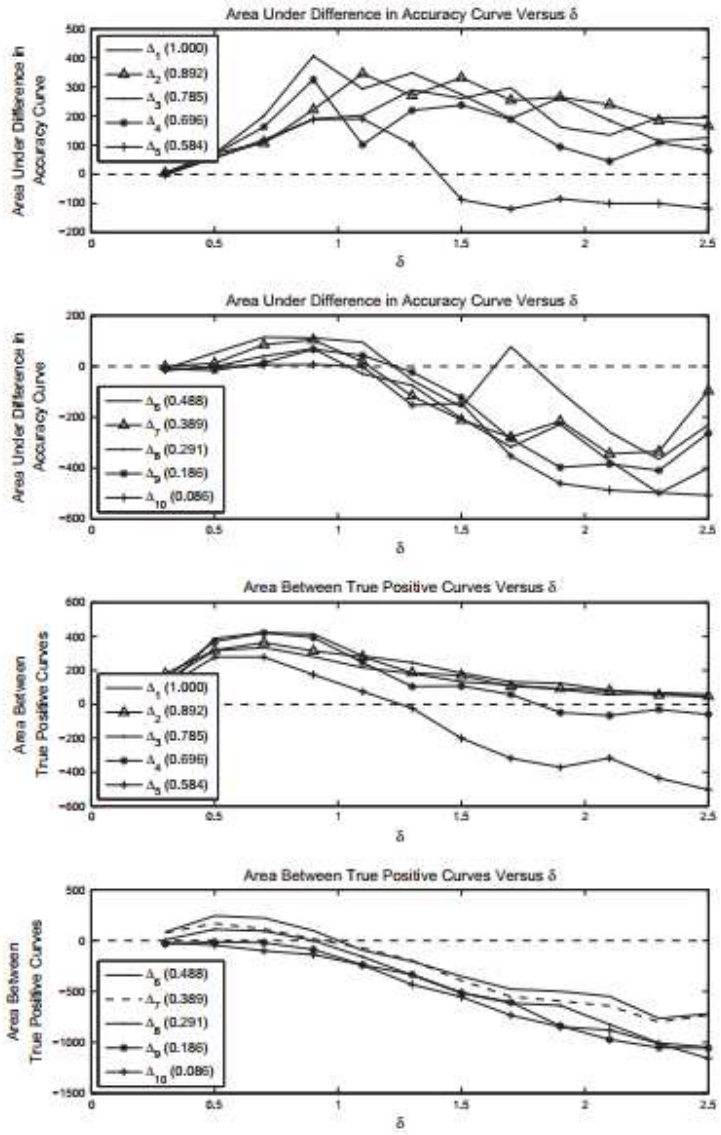
$$\phi_t = |genesCoveredBy(t) - covered|$$
$$covered = covered \cup genesCoveredBy(t)$$

پایان حلقه ی شرطی

مقدار $\frac{\|\phi\|}{|\Delta|}$ را باز گردان

الگوریتم 3 یک الگوریتم پر کاربرد است که حداقل پوشش مجموعه ها از عبارت G را تخمین میزند که از مرجع 52 استفاده شده است. ایده ی کلی آن به طور تصویری در شکل 2 نشان داده شده است. این الگوریتم در ابتدا عبارت t را میابد که بیشترین ژن ها را پوشش میدهد و ϕ_t را به عنوان تعداد ژن ها پوشش داده شده توسط عبارت t در نظر میگیرد. در تکرار بعدی ، الگوریتم عبارت t' را میابد که بیشترین ژن هایی را پوشش میدهد که از قبل پوشش نیافتند و مقدار $\phi_{t'}$ ره به عنوان تعداد ژن هایی که توسط t' پوشش داده میشود در نظر میگیرد. این روند تا زمانی ادامه پیدا میکند که تمام ژن های پوشش یافته و ما ژن ها را دو بار به حساب نیاوریم. این کار این امکان را به ما میدهد تا بردار ϕ را توسط $|\Delta|$. نرمال سازی کنیم. مقدار بهینه ،

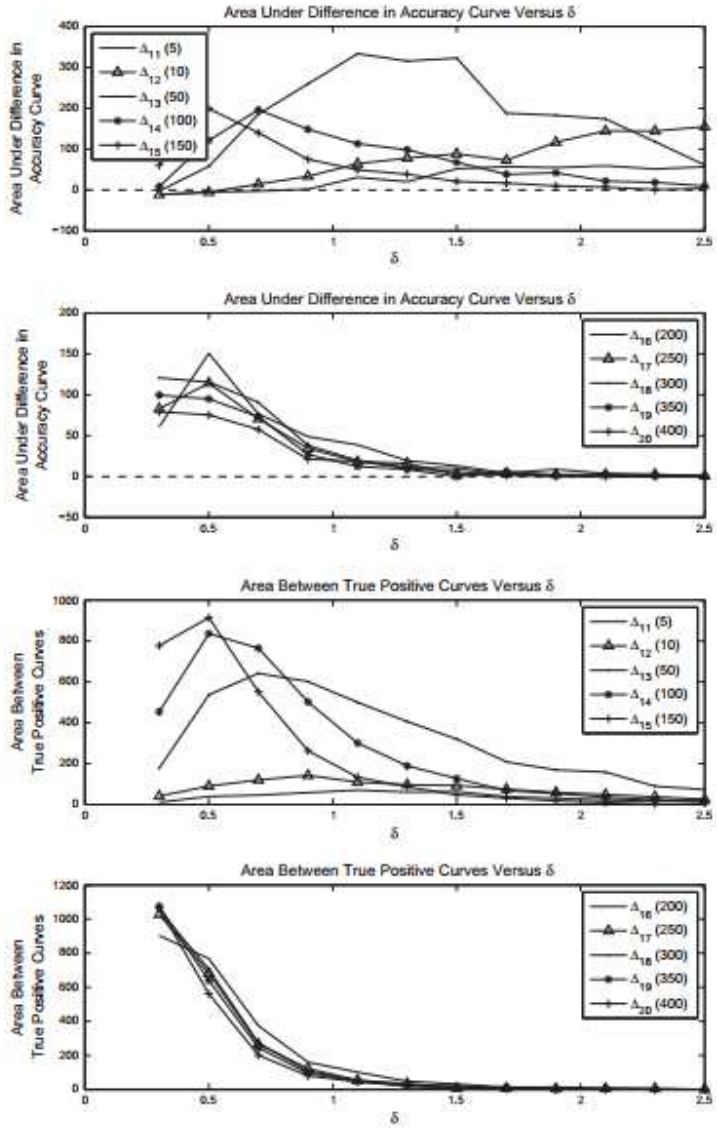
$BCAL(G) = 1$ هنگامی اتفاق میافتد که تمام ژن ها توسط یک عبارت منفرد تفسیر شود ، که گراف ستاره است. بنابراین تمام ژن ها ، g_1 و g_2 ؛ دارای $Sim_{MAX}(g_1, g_2) = 1$ هستند. ازین رو دانستن یک ژن مشخص به الگوریتم 2 این امکان را میدهد تا تمام دیگر ژن های مشخص را از طریق عبارت مشترک استنباط کند. $BCAL(G)$ با تقسیم $|\Delta|$ نرمال سازی میشود تا $0 \leqslant BCAL(G) \leqslant 1$. دلیل برای این کار این است که $\|\phi\| = \sqrt{\sum_{i=1}^{|Terms|}\phi_i^2} \leqslant \sqrt{\left(\sum_{i=1}^{|Terms|}\phi_i\right)^2} = \sum_{i=1}^{|Terms|}\phi_i \leqslant |\Delta|$. تقسیم کردن هر دو طرف نا معادله بر $|\Delta|$ بیان میکند که $\frac{\|\phi\|}{|\Delta|} \leqslant 1$. یک ضرر در $BCAL(G)$ این است که دارای محتویات اطلاعاتی β' است . در بررسی های ما ، ما $\beta' = 1$ در نظر گرفتیم. در صورتی که $\beta' = 1$ ، بنابراین بعضی از حالات زیستی ممکن است $BCAL(G)$ مصنوعی پایینی بگیرند. این ممکن است زمانی رخ بدهد که ژن هایی که یک حالت زیستی را توصیف میکنند عموما از طریق عبارت t که $IC_{norm}(t) < 1$ مرتبط میشوند. اما ، پارامتر β' باید در بیشترین مقدار ممکن باشد ، زیرا شباهت معنایی بین دو عبارت توسط محتویات اطلاعاتی پایین ترین جد مشترک آن دو محدود شده است. ازین رو استفاده از یک β کم ممکن است یک $BCAL(G)$ مصنوعی بالا به دست بیاورد.



شکل6

ما اکنون معنی زیستی $|BCAL(G)|$ را گسترش میدهیم. فرض کنید ما دو گروه از مریض ها را با هم مقایسه میکنیم و 40 ژن درست که به طور متفاوت بیان شده اند بین گروه های مریض های وجود دارد. در صورتی که تمام این ژن ها به یک عبارت منفرد برگیGOتفسیر شوند ، سپس این شرایط زیستی دارای $|BCAL(G) = 1|$. در این حالت ، تمام ژن ها دارای محصولات ژنی هستند که در روند زیستی مشترکی درگیر میباشد. ممکن است روند های زیستی دیگری وجود داشته باشد که جایگزین شده باشد، اما یک روند زیستی مشترک وجود

دارد که بر تمام ژن هایی که به طور متفاوت بیان شده است ، تفسیر ده است. در طرف دیگر طیف ، در صورتی که 40 ژن روی هر کدام از عبارات برگی GO تعریف شده و یا نشده باشند ، و یا هر ژن فقط بر یک عبرت منفرد برگی GO تعریف شده باشد ، سپس این شرایط دارای $BCAL(G) = 0$ خواهد بود. در این حالت ، روند های زیستی وجود دارد که جایگزین شده اما هیچ وجه مشترکی از ژن ها در عبارات برگی GO برای این روند های زیستی وجود ندارد. این حالت بدین معنی است که هر روند زیستی موجود جایگزین در GO حداکثر روی یک ژن تفسیر شده است و در صورتی که شرایط زیستی دارای $BCAL(G) = 0.5$ ، این میتواند توسط داشتن 4 روند زیستی جایگزین شده رخ میدهد که هر کدام بر 10 ژن مختلف تفسیر شده است. در این حالت ، حداقل 4 روند زیستی مختلف وجود دارد که جایگزین شده است و هر کدام از این روند ها دارای 10 ژن است که به طور متفاوت بیان شده است. خصوصا ، اندازه گیری است برای نشان دادن این که ژن ها در عبارات روند های خاص زیستی چقدر به یکدیگر نزدیک هستند. $BCAL(G)$ همچنین در صورتی که ژن های بیان شده به صورت متفاوت در روند های کمتر زیستی درگیر باشند ، نزدیک تر به یک خواهد بود. $BCAL(G)$ در صورتی نزدیک تر به صفر خواهد بود که ژن هایی که به طور متفاوت بیان شده اند دارای وجه مشترک کمی در میان روند های زیستی داشته باشند. یک علت ممکن برای $BCAL(G)$ پایین میتواند عدم وجود اطلاعات کافی در GO برای شرایط زیستی تحت مطالعه باشد.



شکل 7

4. نتایج و بحث ها

ما 20 شرایط زیستی مصنوعی را در نظر میگیریم. ما تحلیل ها را به دو گروه از نتایج تقسیم میکنیم. گروه اول از 10 شرایط زیستی $(\Delta_1, \Delta_2, \ldots, \Delta_{10})$ که ما تحلیل میکنیم و دارای مقادیر $BCAL(G)$ متغیر از 0.086 تا 1 هستند. حالت زیستی بنیادی Δ_{01} با استفاده از الگوریتم 1 ایجاد شده است و ما این حالت زیستی را برا به

وجود آوردن دیگر حالت های اصلاح میکنیم. این حالت زیستی Δ_0 دارای $|\Delta_0| = 36$ و $BCAL(G) = 0.68.$ است.

گروه دوم از 10 حالت زیستی $(\Delta_{11}, \Delta_{12}, \ldots, \Delta_{20})$ همگی دارای $BCAL(G) = 1.000.$ هستند. اما ، این حالت های زیستی دارای قوت های مختلف از 5 ت 400 هستند. هدف تحلیل این گروه از حالت های زیستی یافتن تعداد ژن هایی است که بهبود گزینش مبتنی بر GO را تاثیر میدهند. در این بخش ، ما ابتدا در رابطه با نتایج شرایط زیستی Δ_1 به تفضیل بحث میکنیم و سپس ، ما خلاصه ای از نتایج مرتبط با $\Delta_1, \Delta_2, \ldots, \Delta_{10}$ را مورد بحث قرار میدهیم. سپس ما خلاصه ای از نتایج $(\Delta_{11}, \Delta_{12}, \ldots, \Delta_{20})$ مورد بحث قرار میدهیم.

4.1 شرایط زیستی Δ_1

ما اکنون نتایج برای Δ_1 را تحلیل میکنیم. برای یاد آوری ، این حالت زیستی دارای $BCAL(G) = 1.00$ و $|\Delta_1| = 23$ میباشد. شکل 3 نتایج آزمایش های ما را با $\delta = 0.3$ و شکل 4 نتایج را هنگامی که $\delta = 1.1$ و شکل 5 نتایج را هنگامی که $\delta = 1.9$ و 2.3 است نشان میدهد. هر کدام از شکل ها دارای دو زیر رسم هستند. هر کدام از آن ها مطابق یک مقدار مشخص از δ بوده و سه پنل وجود دارد که داخل آ« ها هذ زیر رسم وجود دارد. پنل بالایی یک وقفه ی 95 ٪ اطمینانی را برای تفاوت در صحت طبقه بندی بین طبقه بندی تمرین داده شده با گروه غنی شده ی ژن ها و طبقه بندی کننده استفاده کننده از گروه تست t ژن ها را نشان میدهد. پنل میانی میانگین صحت طبقه بندی برای هر کدام از طبقه بندی کننده ها را روی 20 تکرار آزمایشی نشان میدهد و نهایتا پنل پایینی تعداد ژن های درست مثبت برای هر گروه ژن را نشان میدهد.

هنگامی که $\delta = 0.3$ ، هیچ بهبودی در صحت طبقه بندی با استفاده از گروه غنی شده وجود ندارد . یک بهبود در شناسایی ژن های درست مثبت برای گزینش مشخصه مبتنی بر GO وجود دارد. برای مثال ، هنگامی که تعداد نمونه های تمرینی برای هر طبقه حدود 100 است ، یک گروه غنی شده دارای میانگینی حدود 5 ژن

درست مثبت است در حالی که گروه تست t دارای میانگین 2 است. هیچ بهبودی در صحت طبقه بندی وجود ندارد زیرا یک $\delta = 0.3$ در متمایز سازی بین طبقه با این تعداد کم از ژن ها نا کافی است.

در حالتی که $\delta = 0.7$ ، یک بهبود در صحت طبقه بندی وجود دارد هنگامی که تعداد نمونه های تمرینی بیشتر از 10 تا در هر طبقه است. اما ، این بهبود بسیار کم است که میانگین حدود 2 ٪ برای تمام سایز نمونه دارد. صحت کلی طبقه بندی نیز حدود 50 تا 60 ٪ است که بسیار کم است. گروه غنی شده شامل ژن های مثبت صحیح بیشتری از گروه تست t برای تمام سایز های نمونه های تمرینی است.

مادامی که δ تا 1.1 افزایش پیدا میکند ، مقدار بهبود نیاز افزایش پیدا میکند. میانگین بهبودی تا حدود 4 ٪ رسیده در زمانی که تعداد نمونه های تمرینی در هر طبقه 40 تا، 80 تا و 100 است. یک الگوی مشابه نیز هنگامی رخ میدهد که δ = 1.5 است. هنگامی که δ افزایش پیدا میکند و به 1.9 میرسد ، بهبودی نزدیک به 5 ٪ در سایز های کوچک نمونه میشود و مادامی که سایز نمونه ی تمرینی افزایش پیدا میکند به صفر میرسد. این حالت مشابه حالتی است که $\delta = 2.3$ است. میزان بهبودی به سمت صفر کاهش پیدا میکند زیرا هر دو گروه غنی شده و تست t در میانگین شامل تقریبا تمام 23 از Δ_7 میشوند.

با بررسی تمام گستره ی نتایج یک الگو ظاهر میشود. ضرورتا ، نتایج یک منحنی نزولی مقعر را دنبال میکنند ، که از سایز نمونه ی بزرگ به سایز های کوچکتر نمونه تغییر پیدا میکند مادامی که δ افزایش پیدا میکند.

این نتایج مشخص میکند که یک افزایش در صحت طبقه بندی تقریبا در تمام موارد وجود دارد هنگامی که $BCAL(G) = 1.000.$ است. اما ، این حالت ایده آل است ، در صورتی که ما $BCAL(G)$ را کاهش دهیم چه رخ میدهد؟

4.2 نتایج برای مقادیر متفاوت $BCAL(G)$

کاملا واضح است هنگامی که $BCAL(G)$ 1 باشد ، تقریبا همیشه استفاده از گزینش مشخصه مبتنی بر GO مفید است. اما ، این حالتی ایده آل را نشان میدهد. کاملا غیر محتمل است که داده های واقعی بیان ژن ها دارای $BCAL(G)$ 1 باشند. در این بخش ، ما حالت های زیستی 1 تا 10 را بررسی میکنیم. ما نمیتوانیم تمام نتایج برای هر حالت زیستی را مانند توضیحات برای Δ_1 ارائه کنیم ، زیرا این نیازمند توضیحات بسیار است. ازین رو ما از روش خلاصه سازی استفاده میکنیم ، که یک مقدار منفرد برای هر رسم را محاسبه میکند. برای درک این مقادیر ، شکل های 5-3 را به یاد بیاورید. ما سه متغیر را با هم مقایسه میکنیم : δ ،سایز نمونه ی تمرینی ، و تفاوت در صحت طبقه بندی . بنابراین برای فشرده سازی این حالت به مقایسه دو پارامتر ، ما ناحیه زیر منحنی صحت طبقه بندی را نسبت به تعداد نمونه تمرینی محاسبه میکنیم. در حالت Δ_1 هنگامی که $\delta = 0.3$ ، ما ناحیه ی زیر منحنی در پنل بالای شکل 3 محاسبه میکنیم. با انجام دادن این یک مقدار منفرد به دست می آید که در هر مقدار δ برای شرایط زیستی است. در صورتی که مساحت زیر منحنی مثبت باشد ، یک بهبودی خالص در استفاده از گزینش مشخصه مبتنی بر GO وجود دارد. در صورتی که ناحیه ی زیر منحنی منفی باشد نیز هیچ بهبودی وجود ندارد. در واقع ، در صورتی که ناحیه ی زیر منحنی منفی باشد ، سپس استفاده از روش گزینش مشخصه مبتنی بر GO موجب یک کاهش خالص در صحت طبقه بندی در گستره ی نمونه ی تمرینی میشود. فواید استفاده از این روش خلاصه سازی این است که این امکان را به ما میدهد تا تمایل کلی تاثیر $BCAL(G)$ بر روی صحت طبقه بندی را ببینیم. ازین رو به نوبه ی خود یک بصری سازی کوتاه نسبت به تمام نتایج را به دست میدهد. بدی آن نیز این است که ما بعضی از جزئیات را نسبت به زمانی که روش ما موجب افزایش صحت میشود ، از دست میدهیم. برای مثال ، گزینش مشخصه مبتنی بر GO میتواند موجب افزایش صحت طبقه بندی در سایز کوچک نمونه شود سپس در سایز های بزرگ میزان صحت کاهش پیدا کند. این اطلاعات هنگامی که ما ناحیه ی زیر منحنی را محاسبه میکنیم گم میشود. به علاوه ی

ناحیه ی زیر قسمت تفاوت در منحنی صحت ، ما همچنین ناحیه ی بین گروه غنی شده و گروه T منحنی های مثبت را نیز محاسبه میکنیم.

شکل 6 خلاصه از ار مساحت زیر تفاوت در منحنی صحت و ناحیه ی بین منحنی مثبت درست گروه غنی شده و گروه t را نشان میدهد . هنگامی که $BCAL(G) \geqslant 0.69$ است ، یک بهبودی خالص در صحت طبقه بندی برای طبقه بندی کننده ای که از گروه غنی شده استفاده میکند ، وجود دارد. هنگامی که $\delta = 0.3$ است هیچ بهبودی در صحت وجود ندارد. گزینش مشخصه ی مبتنی بر GO تا زمانی که $0.9 \leqslant \delta \leqslant 1.1$ موجب بهبودی میشود که میزان بهبودی به حداکثر میرسد. بعد از این میزان حداکثر ، سطح بهبودی های هنگامی که $BCAL(G) \geqslant 0.696$ است کاهش پیدا میکند. در موردی که $BCAL(G) < 0.696$ این روش شروع به کاهش در صحت طبقه بندی میکند که $\delta \geqslant 1.5$ است. جالب است بدانیم که هر دو حالت زیستی Δ_2 و Δ_3 نسبت به Δ_1 عملکرد بهتری دارند حتی با وجود این که این دو حالت زیستی دارای مقادیر کمتر $BCAL(G)$ هستند. ما بر این باور هستیم که این به علت این حقیقت است که سایز گروه ها 40 است و $|\Delta_1| = 23$ ، $|\Delta_2| = 26$ و $|\Delta_3| = 31$ میباشد و آن ها در سایز به تعداد ژن های انتخاب شده نزدیک تر هستند. بنابراین با Δ_1 ، الگوریتم 2 از ژن های مثبت صحیح خالی میشود و نمیتواند چیزی به گروه غنی شده اضافه کند. یک مشاهده ی جالب دیگر این است که هنگامی که $\delta > 1$ است ، حالت زیستی با $BCAL(G) = 0.696$ (Δ_4) مثبت های صحیح بیشتری نسبت به Δ_1 و Δ_2 به ست می آورد. ما باور داریم که این نیز در اثر قوت Δ_4 ایجاد شده است که 35 است.

نمودار مثبت صحیح شکل 6 ایده های مهمی را نشان میدهد. در مواردی که $BCAL(G) \geqslant 0.584$ ، تمایل کلی برای ناحیه ی بین منحنی های مثبت صحیح این است که آن ها هنگامی که $\delta = 0.3$ است با مقدار کمی شروع میشوند ، و آن ها زمانی که $\delta = 0.7$ میشود به بیشترین مقدار خودشان میرسند. منحنی ها

سپس مادامی که δ افزایش پیدا میکنند شروع به رفتن به سمت صفر میکنند. اما ، برای هر دو حالتی که $BCAL(G) = 0.696$ و $BCAL(G) = 0.584$. است ، گروه غنی شده نسبت به گروه تست t مقادیر کمتری از مثبت های صحیح را به دست می آورد. Δ_4 در $\delta = 1.7$ به زیر صفر رفته و Δ_5 در $\delta = 1.3$ به زیر صفر میرسد. هنگامی که $BCAL(G) \leqslant 0.488$ منحنی به مقدار پیک خودش رسیده و حتی زودتر به این مقدار میرسد. در حالتی که $BCAL(G) = 0.086$. است ، مقادیر پیک در $\delta = 0.3$ رخ میدهد. در تمام حالت ها هنگامی که $BCAL(G) \leqslant 0.488$ است منحنی ها قبل از $\delta = 1$ به زیر صفر میروند. در بدترین حالت ، ناحیه ی بین گروه غنی شده و گروه تست t نزدیک به 1000- است. مقدار این بیش از دوبرابر بهترین حالت است که حدود 400 است. این بیان دارد که گزینش مشخصه ی مبتنی بر GO میتواند به طور محسوسی بد تر از فواید بالقوه اش در زمانی که $BCAL(G) \leqslant 0.291$ و $\delta \geqslant 2.3$. است باشد.

4.3 نتایج برای مقادیر مشابه $BCAL(G)$

یک سوال مهم که ما آن را بررسی نکرده ایم این است : با در نظر داشتن یک $BCAL(G)$ ثابت ، تعداد ژن ها چگونه میتواند تفاوت در صحت طبقه بندی را تاثیر دهد ؟ برای پاسخ دادن به این سوال ، ما $BCAL(G) = 1.000$ را ثابت کرده و تعداد ژن ها را تغییر میدهیم. ما 10 حالت زیستی $\Delta_{11}, \Delta_{12}, \ldots, \Delta_{20}$ را با تعداد ژن هایی که از 5 تا 10 و سپس از 50 تا 400 ژن که در هر بار 50 تا اضافه میشود ایجاد میکنیم. این به ما امکان میدهد تا کارایی روش گزینش مشخصه GO را هنگامی که تعداد ژن ها در شرایط زیستی ناجور میشود که یا خیلی بزرگ و یا خیلی کوچک میشود را ارزیابی کنیم. ما $BCAL(G) = 1$ ثابت کردیم زیرا ما میتوانیم در این حالت تعداد ژن ها را بسیار بهتر کنترل کنیم. ما همچنین تعداد ژن های انتخاب شده برای طبقه بندی را روی 40 نیز تثبیت کردیم ، مشابه حالت قبل. شکل 7 نتایج این نتایج شبیه سازی را نشان میدهد.

تمایل هایی در شکل 7 وجود دارد که مهم است در رابطه با آن ها بحث کنیم. در دو منحنی بالایی ، ما ناحیه ی زیر تفاوت ها در صحت طبقه بندی را میبینیم. مشابه شکل 6 ، مقدار مثبت نشان دهنده ی بهبود در گروه غنی شده نسبت به گروه تست t و مقدار منفی نشان دهنده کاهش صحت است. گروه غنی شده ضرورتا بهبودی را برای Δ_{11} ایجاد نمیکند. برای حالت Δ_{12} گروه غنی شده هنگامی که $\delta > 1$ است بهبود ایجاد میکند و هنگامی که $\delta \leqslant 1$ است بهبودی بسیار کم است و یا وجود ندارد.شرایط زیستی Δ_{13} دارای شکلی مشابه به Δ_1 از شکل 6 است. هنگامی که تعداد ژن ها تا 100 تا در Δ_{14} افزایش پیدا میکند مقدار بهبود ها به طور محسوسی کاهش پیدا میکند. این بیان دارد که تست t میتواند مثبت های صحیح بسیار بیشتری را با Δ_{14} در مقایسه با Δ_{13} به دست بیاورد. اما ، زمانی که $\delta < 0.7$ است ، گروه غنی شده عملکرد بهتری را نسبت به گروه تست t به دست آورد زمانی که تعداد ژن ها در شرایط زیستی حداقل 150 تا است. در صورتی که ما دو منحنی پایینی شکل 7 را بررسی کنیم ، ما این شرایط را بسیار بهتر درک میکنیم. هنگامی که $\delta = 0.3$ ، مادامی که تعداد ژن ها در شرایط زیستی افزایش پیدا میکند ، گوره غنی شده ژن های مثبت صحیح بیشتری دارد. این بدین علت رخ میدهد که شانس بیشتری برای پیدا شدن ژن های مثبت صحیح در گروه ژن های رتبه بندی شده اولیه وجود دارد. این موضوع به الگوریتم 2 این امکان را میدهد تا ژن های مشابه معنایی بسیاری را به ژن های گروه غنی شده اضافه کند. این بهبودی مادامی که ژن ها در شرایط زیستی به 250 میرسند به سمت همگرایی پیش میرود. هنگامی که $\delta > 1$ و تعداد ژن ها در حالت زیستی بیشتر از 150 است ، استفاده از روش گزینش مشخصه مبتنی بر GO فایده ای کم داشته و یا بی فایده است. ازین رو ، ما استفاده از این روش را برای شرایط زیستی که ژن هایی بیش از 150 تا و $\delta > 1$ و $BCAL(G) < 1$ پیشنهاد نمیکنیم.

5. خلاصه و نتایج

این مطالعه ی به بررسی این سوال میپردازد که چه زمانی باید از روش گزینش مشخصه ی مبتنی بر GO به طور موثر استفاده کرد ، در حالی که مطالعه های قبلی روش های را برای چگونگی استفاده از این روش ارائه کرده اند. برای بررسی این سوال ، ما یک شبیه سازی را ایجاد کردیم. گام اول شبیه سازی الگوریتم 1 است که مجموعه ای از ژن ها را به عنوان خروجی تحویل میدهد. این ژن ها به طور متفاوتی بین یک طبقه کنترلی و یک طبقه ی آزمایشی بیان شده اند. این مجموعه از ژن ها نشان دهنده ی یک حالت زیستی است که با Δ نشان داده میشود. ما داده های بیان ژنی مصنوعی را با استفاده از داده های واقعی جمع آوری شده از سلول های پوششی مسیر هوایی به دست آورده ایم. داده ها از طبقه ی آزمایشی (نشان دهنده ی شرایط زیستی تغییر یافته) بر گروه کنترلی مبتنی است ، به غیر از ژن هایی که در مجموعه ی Δ هستند و بیانشان افزایش و یا کاهش یافته است. دامنه ی افزایش و یا کاهش ها در Δ توسط پارامتر δi کنترل میشود. دو مجموعه داده ها شامل حالت های بنیادی روند های تولید داده هستند که نمونه هایی جدید با نویز جدید را ایجاد میکند. ما دوطبقه بندی کنندهی DLDA را روی داده های تولید شده از این ژن های اولیه تمرین میدهیم. یک طبقه بندی کننده از ویژگی های آماری برای انتخاب ژن ها استفاده میکند و طبقه بندی کننده ی دیگر از ویژگی های آماری در ارتباط با شباهت های معنایی در GO برای گزینش ژن ها استفاده میکند. ما یک اندازه گیری را به نام $BCAL(G)$ معرفی میکنیم که سطح تفسیر را کمی سازی کرده و یا همبندی ژن ها و عبارات را در یک حالت زیستی نشان میدهد.

ما پنج نتیجه ی اصلی را از شبیه سازی هایمان به دست آورده ایم. اولا ، استفاده از گزینش مشخصه مبتنی بر GO زمانی که $BCAL(G) \geqslant 0.696$ سودمند است. $BCAL(G)$ از لیستی از ژن های پتانسیل محاسبه شود. این لیت پتانسیل ممکن است با جستوجو در مقاله ها یافت شود. در عمل ، در صورتی که زیست شناس های انتظار دارند که ژن هایی که به طور متفاوت بیان شده اند احتمالا با تعداد کم عبارات GO مرتبط باشند ($BCAL(G)$ نزدیک به 1) ، احتمالا استفاده از روش گزینش GO باعث بهبود طبقه بندی ها بشود. دوما ، هنگامی که

$BCAL(G) \leqslant 0.389\ (\Delta_7 - \Delta_{10})$ ، گزینش ویژگی های آماری از روش گزینش GO بهتر عمل میکند به جز مورد غیر رایج که در شکل 6 ارائه شده است که در منحنی دوم از بالا $0.5 \leqslant \delta \leqslant 1.2$ است. عملا ، در صورتی که زیست شناس انتظار دارد که ژن هایی که به طور متفاوت بیان شده اند رابطه ی نزدیکی درGO نداشته باشند ($BCAL(G)$ نزدیک به صفر) استفاده از روش GO پیشنهاد نمیشود. سوما ، هنگامی که $0.389 < BCAL(G) < 0.696$ ، و $\delta < 1$ باشد روش گزینش GO بهبودی بین 0 تا 9 ٪ را فراهم میکند. اما ، میانگین بهبود روی تمام سایز های نمونه های تمرینی از 0 تا 2 ٪ است. در عمل در صورتی که لیست پتانسیل ژن های در دو حالت قبلی قرار نگیرد ، پس در این گروه است. اما ، گزینش بر مبنای GO در این حالت تنها در صورتی مفید است که میانگین تغییر در بین ژن ها کمتر از یک واحد در فضای نرمال شده ی RMA باشد.چهارما ، در صورتی که $BCAL(G) = 1$ ثابت شده باشد و ما تعداد ژن ها را در Δ بیش از 50 تا افزایش داده و $\delta \geqslant 0.7$ باشد ، سپس بهبود در روش GO کاهش پیدا میکند. این جا منظور ما این است که میانگین بهبودی با مقدار بزرگتری شروع شده و مادامی که تعداد ژن های مهم افزایش پیدا میکند ، مقدارش کم میشود. این بر این فرض است که سایز گزینش مشخصه ی گروه ثابت است. به طور خاص ، هنگامی که تعداد ژن ها در حالت زیستی بیشتر از 150 است ، و $\delta > 1$ ، ما استفاده از روش GO را پیشنهاد نمیکنیم. عملا ، یک زیست شناس نباید از GO برای بهبود گزیشن مشخصه با ژن های پتانسیل بیش از 150 استفاده کند. پنجما ، ما استفاده از GOرا در صورتی که تعداد ژن ها در شرایط زیستی کمتر از 10 تا است نیز پیشنهاد نمیکنیم.

در حالی که شبیه سازی های ما یک درک از روش گزینش مشخصه ی GO فراهم میکند ، محدودیت هایی نیز دارد. داده های مصنوعی ما از داده های واقعی به دست آمده بود که توسط RMA پیش پردازش شده بود. ازین رو ما باور داریم که نتایج ما برای مجموعه داده هایی معتبر هستند که با استفاده از RMA پیش پردازش شده اند. به جز RMA ، مقادیر δ ممکن است برای روش های مختلف پیش پردازش متمفاوت باشد. دیگر مجدودیت شبیه سازی ما از این حقیقت سر چشمه میگیرد که ما تحلیل های خودمان را به اندازه گیری های

شباهت های معنایی منفرد ، روش های گزینش مشخصه و روش های طبقه بنید محدود کردیم . اما ، ممکن است روش GO بهبود هایی را روی گستره های مشابه $BCAL(G)$ برای اندازه گیری های متفاوت شباهت معنایی ، گزینش مشخصه و روش های طبقه بندی به دست بیاورد. تحقیق های آینده این مشکل را مرتفع خواهد کرد.

به طور کل شبیه سازی ما دیدگاهی نسبت به روش GO ایجاد میکند. مخصوصا شبیه سازی ما نشان میدهد که چه زمانی استفاده از این روش مفید است. ما باور داریم که شبیه سازی ما میتواند به محققین کمک کند تا از این روش GO به طور مفید تر استفاده کنند. به علاوه ، شبیه سازی ما میتواند به طور مثبت بر ابزار تحلیل های غنی سازی و تحلیل های تفسیری بنیادی تاثیر بگذارد زیرا به محققان این امکان را میدهد تا شرایط زیستی را از GO مشخص کرده و کارایی این الگوریتم ها را در شناسایی شرایط خاص زیستی مقایسه کنند. این مقایسه میتواند تاثیر مثبتی بر پروژه هایی مانند نقشه های همبندی داشته باشد زیرا میتواند موجب سطح تفسیر بهبود یافته در نشان دادن شرایط زیستی بشود. علاوه بر ایت ، $BCAL(G)$ میتواند ابزار تحلیلی غنش شده تک ژنی را بهبود دهد زیرا میتواند برای یافتن گروه هایی از عبارات که بهترین پوشش روی لیست ژن مربوطه را ارائه میدهد در سر حد های مختلف محتویات اطلاعاتی بیابد.