کشف مسیرهای تاثیرگذاری ژنها بر یکدیگر بر اساس گروههای ژنی تصحیح شده توسط SVM

امیر اسدی [1] ، حسام دشتی [2,*,] ، نسیم جمالی [3] ، تینگ تینگ ژانگ [4] ، ماری کلاک [5] ، تانگ لی [6] ، گرگوری میشلوتی [7]**

۱. دانشکده ریاضی، دانشگاه مدیسون-ویسکانسین، ایالات متحده آمریکا.
۲. مرکز عالی بیومتمیک، دانشگاه تهران، ایران.
۳. مرکز عالی بیومتمیک، دانشگاه تهران، ایران.
۴. دانشکده تحقیقات مشترک، دانشگاه روچ پالو، ایالات متحده آمریکا.
۵. مرکز تحقیقات بیوفیزیک، دانشگاه مدیسون-ویسکانسین، ایالات متحده آمریکا.
۶. دانشکده روانپزشکی، دانشگاه دوک و مرکز درمانی دورهام، ایالات متحده آمریکا.
۷. دانشکده بی هوشی، دانشگاه دوک و مرکز درمانی دورهام، ایالات متحده آمریکا.

چکیده

در این مقاله، متدی بر اساس ادغام متدهای یادگیری ماشین[1] و متدهای حل معادلات دیفرانسیل برای دستیابی به متدی با دقت بالا جهت کشف مسیرهای تاثیرگذاری ژنها[2] ارائه گردیده است. ما با بکار بردن ابزار Support Vector Machines (SVM) بر متد InfoMax [۱۲]، فرانید گروهبندی[3] این متد را بهینه کرده ایم که الگوریتم ارائه شده در این مقاله حاصل این فرایند بهینه سازی است. فرایند بهینه سازی به اینصورت می باشد که تعدادی از اعضای هر گروه‌ی تولید شده در InfoMax بعنوان مجموعه آموزشی[4] SVM انتخاب می گردند. در طی فرایند یادگیری SVM، گروهبندی اعضای مجموعه آموزشی تغییر می کند. با بکار بردن SVM آموزش داده شده بر روی بیان ژنهای پایگاه داده ای که InfoMax آنرا گروهبندی کرده بود شاهد تغییرات بسیاری بودیم و لذا گروههای تولید شده توسط متد حاضر با گروهبندیهای InfoMax، SOM، و پایگاه داده ای کگ[5][۱۳] مقایسه گردیدند. فرایند مقایسه بر روی پایگاه داده ای از بیان ژنهای موش [۱۲] و خمیر مایه [۱] امتحان گردید. نتایج عالی این مقایسه در بخش نتایج آمده است.

کلمات کلیدی: متدهای گروهبندی، متدهای طبقه بندی[6]، بیان ژنها،مسیرهای تاثیرگذاری ژنها، Support Vector Machines.

مقدمه

آنالیز رفتار سلول های موجودات زنده به منظور ادراک فرآیندهای حیاتی سلول یکی از مهم ترین مسائل مطرح در زمینه تحقیقاتی بیوانفورماتیک به شمار می‌رود. از آنجا که رفتارهای گوناگون سلول از طریق تغییر مقدار کپی‌برداری از ژن هایش -که به آن بیان ژن می‌گوییم [۵]- صورت می پذیرد، آنالیز رفتار سلول از طریق بررسی بیان ژن های سلول صورت می‌گیرد [۶]. در مورد تغییر مقدار بیان ژن، نکته حائز اهمیت، تأثیر تغییر بیان یک ژن بر مقدار بیان سایر ژن هاست. تحقیقات زیست -شناسان در این زمینه، نشان داده است که این تأثیرگذاری از گروهی از ژنها به گروههای دیگری ژنی صورت می‌پذیرد. گروه بندی ژن هایی که تأثیر یکسان بر سایر ژن ها دارند یا به عبارت دیگر هم رفتاراند، از طریق بررسی مکاشفه‌ای میزان تغییر بیان هر یک از ژن ها صورت می‌گیرد. از آنجا که تأثیر بیان ژن ها بر یکدیگر نیازمند گذر زمان است، بررسی تغییر بیان ژن ها بر یکدیگر، نیازمند آنالیز یک سری زمانی از مقدار بیان ژن ها می باشد. از اینروی در متدهای ارائه شده در این زمینه همواره پایگاه داده ای از سلسله بیانهای مختلف ژنها مورد بررسی قرار می گیرد. تا کنون متدهای مختلفی [۲][۳][۴] جهت گروهبندی داده های ژنی ارائه گردیده اند که یکی از قویترین آنها متد ژیااو[7] [۴] می باشد. متد محاسباتی ما بر پایه Support Vector Machines به بهینه سازی گروه بندی ژن های هم رفتار InfoMax می‌پردازد. این متد بر روی پایگاه داده‌ای موش و خمیر مایه آزمایش گردیده است. پایگاه داده ای موش با مقدار بیان ۲۷۰۰۰ ژن در۳ حالت مختلف سلول به دست آمده و پایگاه داده ای خمیر مایه نیز بیان ۶۵۰۰ ژن را نشان می دهد. بیان ژن های موش، از سلول های مبتلا به بیماری پارکینسون جمع آوری شده است. در راستای مقایسه متدها، فرایند تولید مسیرهای تاثیرگذاری که در InfoMax ارائه گردید بر روی گروههای بهینه شده ما نیز اعمال گردید. تطبیق بیشتر این مسیرها با مسیرهای موجود در پایگاه داده ای کِگ بیانگر افزایش دقت محاسبات، و مثبت بودن فرایند بهینه سازی است.

مواد و روش ها

قبل از ارائه متد طراحی شده در ابتدا سه الگوریتم استفاده شده در این مقاله را معرفی می کنیم و سپس به بررسی روند بکار بردن این الگوریتم ها بر یک پایگاه داده ای خاص می پردازیم. بدیهی است می توان متد را برای آنالیز هر پایگاه داده ای بیان ژنی استفاده کرد و صرفا جهت ارائه توضیح شفاف از یک مثال خاص استفاده گردیده است.

از آنجا که این مقاله بر پایه استفاده از ابزار SVM بنا نهاده شده، قبل از ارائه فرایند محاسبه به معرفی کوتاهی بر SVM می پردازیم.

Support Vector Machines (SVM)

SVM یک نوع الگوریتم یادگیری ماشین[8] است که توسط وپنیک[9] و بوزر[10] [۹][۸][۷]، معرفی شد. این نوع از الگوریتم های یادگیری، توانایی بسیاری در آنالیز داده های خطا دار دارند و لذا در زمینه های مختلف، از دسته بندی تصاویر گرفته تا یافتن

---

[1] Machine Learning
[2] Gene pathways
[3] Clustering
[4] Training set
[5] KEGG
[6] Classification
[7] Xiao. Xiang
[8] Machine Learning

سرطان به کار برده می شوند. این گونه از الگوریتم های یادگیری ماشین، جهت ارائه متدی سریع و با دقت بالا در طبقه بندی داده ها[11] تولید و گسترش داده شده اند. SVM با استفاده از پردازش قبلی داده ها ($X \in R^d$) تابعی ($F(X) = W * X + b'$) جهت تمایز گروه ها ارائه می دهد که می تواند به آسانی داده های جدید را طبقه بندی کند. با آنالیز کردن مجموعه داده هایی که جهت یادگیری ماشین داده شده اند، داده های مرزی که بیشترین مرز را بین کلاس ها ایجاد می کنند، مشخص می گردند. و زمانی که متغیر های تابع *F(X)* (*b* و *W*) از روی نمونه های مرزی تعیین شوند، معادله‌ی *F(X)* به یک معادله‌ی درجه دو تبدیل می شود و آماده به کار بستن می شود. بر اساس توزیع داده ها در فضا، SVM ها به دو نوع خطی و غیر خطی تقسیم می شوند. SVM خطی در شرایطی به کار می رود که داده هایی که جهت آموزش SVM به کار برده شده اند، به وسیله ی *F(X)* در فضای کنونی داده ها تفکیک -پذیرباشند. در SVM غیر خطی این داده ها باید به فضایی با بعد بالاتر منتقل شوند تا بتوان به وسیله *F(X)* آن ها را تفکیک نمود. در SVM غیر خطی بر اساس چگونگی توزیع داده ها می توان توابع انتقال ($\Phi : R^d \rightarrow H$) مختلفی استفاده کرد. معمولاً تعریف و استفاده از تابع انتقال مشکل می باشد و لذا در متد SVM تابع هسته[12] تعریف می شود. تابع هسته، عموماً تابع غیر خطی است و به SVM این امکان را می دهد که بجای انتقال داده ها به فضایی با بعد بیشتر و آنالیز آن ها در آن فضا، نتیجه این آنالیز را با اعمال تابع هسته بر داده ها در فضای کنونی به دست آورد [۱۰]. تابع هسته نقش مهمی در تعیین متغیر های تابع *F(X)* در فضای داده ها دارد، به طوری که استفاده از تابع هسته مناسب باعث می شود نمونه های غیر خطی بتوانند به صورت خطی جدا پذیر شوند[۱۱].

متد

همان گونه که بیان گردید برای محاسبه شبکه تأثیر گذاری ژن ها بر یکدیگر باید روند تغییر بیان ژن ها در یک سری زمانی مورد بررسی قرار گیرد؛ از این روی گروه بندی ژن ها نیز بر اساس روند تغییر بیان ژن در سری زمانی صورت می پذیرد. پایگاه داده ای بیان ژن های موش که در ابتدای مقاله معرفی شد، شامل ۱۱ سری بیان ژن می باشد که ۵ سری از این بیان ژنها به شرایطی که داروهای پرگولاید[13] و کینسترین[14] اعمال گردیده اند تعلق دارند و ۵ سری دیگر آن مربوط به بیان ژن ها با اعمال دارو پرگولاید می باشند و ۱ سری باقی مانده از این داده ها به سلول سالم تعلق دارند. بیان ژن سلول سالم جهت بررسی روند بهبود سلول با اعمال داروهای فوق در فرآیند آنالیز وارد گردیده است. جهت همسان سازی این سری داده ها، از داده هایی که به یک حالت تعلق دارند، میانگین گرفته شده و آنالیز بر روی مقادیر میانگین صورت می گیرد. و لذا بجای ۱۱ سری بیان ژن، صرفاً ۳ سری بیان ژن مورد بررسی قرار می گیرند. در نتیجه متد با پایگاه داده ای از بیان ژن ها (E1, E2, E3)، و نسبت بین آن ها (R1=E2/E1,R2=E3/E2) مواجه است و گروه بندی ژن ها را بر اساس آنالیز این نسبت ها انجام می دهد. پس از اعمال متد گروهبندی ارائه شده در InfoMax، ۱۰% از عناصر هر گروه به عنوان مجموعه آموزشی به متد SVM داده می شود، در روند تکرار آموزش SVM برخی از گروه های تعیین شده به وسیله InfoMax تغییر کرده و گروه بندی بهینه می گردد، جدول ۱ روند این بهینه سازی را در تکرارهای مختلف آموزش SVM نشان می دهد. سپس با استفاده از SVM تعلیم داده شده، باقی ژن ها که از آن ها در تعلیم SVM استفاده نشده بود، گروه بندی می گردند.

جدول ۱. روند تأثیر آموزش متد SVM بر گروه بندی داده های ژنی.

Epoch #	Percent of modifications
10	0
50	2
100	18

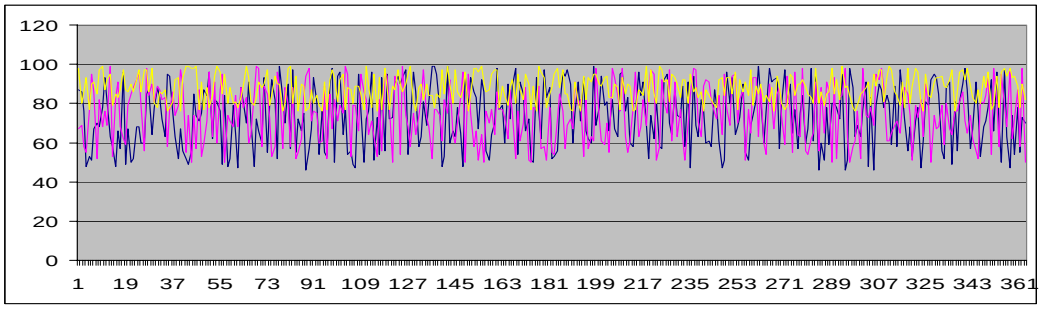
هر ژن دارای دو مقدار R1 و R2 می باشد و لذا می توان هر ژن را برداری در فضای دو بعدی R1 و R2 در نظر گرفت. این ایده موجب شد تا تابع هسته به کار برده شده در این آنالیز، ضرب داخلی بردارهای ژنی باشد. از آنجا که در فرآیند طبقه بندی ژن ها، هدف تشخیص فاصله هر ژن با گروه های ژنی از پیش تعریف شده است و این گروه ها بر اساس تغییر نسبت بیان ژن ها (R1, R2) به وجود آمده اند، به کار بردن تابع ضرب داخلی، تابع خوبی برای طبقه بندی داده ها به شمار می رود.

نتایج

با توجه به آنچه در قسمت های گذشته درباره روند بهینه سازی متد InfoMax بیان گردید، در این فصل به مقایسه گروههای بدست آمده از InfoMax، SOM، و کگ با گروههای حاصل از متد حاضر بر پایگاه داده ای موش می پردازیم. در ابتدا گروهبندی موجود برای ژنهای پایگاه داده را از کگ استخراج نمودیم و سپس متدهای فوق را بر پایگاه داده اعمال کردیم. سپس میزان آشفتگی[15][۱۲] هر یک از گروهها محاسبه گردید و گروههای هر متد بر اساس این میزان آشفتگی مرتب گردیدند. شکل ۱. مقدار شباهت متد کگ [رنگ زرد]، متد SOM [رنگ صورتی]، و متد InfoMax[رنگ آبی] را با متد ما نشان می دهد.

---

9 Vapnik
10 Boser, et al.
11 Classification
12 Kernel Function
13 Pergolide
14 Kinesterin
15 Entropy



شکل ۱. نمودار مقایسه گروهبندی متد ها.

منابع

1. Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W. (1998) "A Genomewide Transcriptional Analysis of the Mitotic Cell Cycle," Molecular Cell, Vol. 2, pp. 65-73.
2. Yaron Hakak, John R. Walker, Cheng Li, Wing Hung Wong, Kenneth L. Davis, Joseph D. Buxbaum, Vahram Haroutunian, and Allen A. Fienberg, (2001) Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia, Proc Natl Acad Sci U S A. 2001 Apr 10;98(8):4746-51.
3. Amir Ben, Ron Shamiry, Zohar Yakhiniz, (1999), Clustering Gene Expression Patterns, J Comput Biol. 6(3-4):281-97.
4. Xiang Xiao, Ernst R. Dow, Russell Eberhart, Zina Ben Miled, Robert J. Oppelt, (2003) Gene Clustering Using Self-Organizing Maps and Particle Swarm Optimization, International Parallel and Distributed Processing Symposium.
5. Genes and Gene Expression, The Virtual Library of Biochemistry, Moleculer Biology and Cell Biology, http://www.biochemweb.org/genes.shtml.
6. Stefan Lorkowski (2002), Analyzing Gene Expression: A Handbook of Methods, Possibilities and Pitfalls, Wiley-VCH
7. B. E. Boser; I. M. Guyon; V .Vapnik (1992), "A training algorithm for optimal margin classifiers"; In Fifth Annual Workshop on Computational Learning Theory, pages 144–152, Pittsburgh, ACM.
8. B. Sch¨olkopf; C. Burges; V. Vapnik (1995), "Extracting support data for a given task"; In U. M. Fayyad and R. Uthurusamy, editors, Proceedings, First International Conference on Knowledge Discovery & Data Mining. AAAI Press, Menlo Park, CA.
9. V. Vapnik (1995), "The Nature of Statistical Learning Theory"; Springer-Verlag, New York.
10. Nello Cristianini; John Shawe-Taylor (2000), "An Introduction to Support Vector Machines and Other Kernel-based Learning Methods".
11. Richard O.Duda; Peter E.Hart; David G.Stork (2003), "Pattern Classification"; Elsevier Academic Press; 2nd Edition.
12. Hesam T. Dashti, Mary Kloc , Tong Lee, Gregory Michelotti, Tingting Zhang, Amir Assadi (2007), InfoMax Gene Networks Constructed from Intervention in the Animal Models of the Parkinson Disease, Sixteenth Annual Computational Neuroscience Meeting.
13. KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/

Discovering genes' pathways based on improved genes' clusters via Support Vector Machines

Amir Assadi [1], Hesam Dashti [*,,2], Nasim Jamali [3], Mary Kloc [4], Tong Lee [5], Gregory Michelotti [6], Tingting Zhang [7]**

1. Department of Mathematics, University of Wisconsin-Madison, Madison, WI 53706, USA.
2. Center of Excellence of Biomathematics, University of Tehran, Iran.
3. Center of Excellence of Biomathematics, University of Tehran, Iran.
4. Biophysics Program, University of Wisconsin-Madison, Madison, WI 53706, USA.
5. Department of Psychiatry, Duke University & Medical Center, Durham, NC 27710, USA.
6. Department of Anesthesiology, Duke University & Medical Center, Durham, NC 27710, USA.
7. Department of Research Partnering, Roche Palo Alto LLC, CA 94304, USA.

We develop an integrated method of machine learning and differential equation methods to reach high accuracy method for discovering genes' pathways. Our method is an improvement of Dashti et. al [12] method, where Support Vector Machines (SVM) are applied for refining this method's gene clustering process. We arbitrarily select percents of InfoMax [12] clusters' members as SVM's training samples. During the training epochs SVM refine the training samples and so some clusters' members changed. Applying trained SVM on whole genes in the data set modifies clustering result and so on pathway detection result. The brilliant result of generated gene pathways shows the affects of using SVM to modify clustering process. The experiment performed on evaluating clustering ability of proposed method by origin InfoMax [12] and SOM [4] methods. These comparisons performed on rat [12] and yeast [1] gene expressions.

Keywords: Clustering methods, Classification methods, Gene expressions, Gene Pathways, Support Vector Machines.

WWW.IBP.IR iranian bioinformatics portal