Iranian Journal of Biomedical Engineering
11 (3) (2017) 219-230, www.ijbme.org
DOI: 10.22041/ijbme.2018.75885.1294

Dimensionality Reduction of Binding Site Sequence Data on Human Genome Using a Deep Autoencoder Neural Network

H. Bankikoshki[1], S. A. Seyyedsalehi[2*], F Zareh-Mirakabad[3]

[1]MSc Student, Bioelectric Department, Biomedical Engineering Faculty, Amirkabir University of Technology, Tehran, Iran
[2]Associate Professor, Bioelectric Department, Biomedical Engineering Faculty, Amirkabir University of Technology, Tehran, Iran
[3]Assistant Professor, Faculty of Mathematics & Computer Sciences, Amirkabir University of Technology, Tehran, Iran

Receipt in the Online Submission System: 20 November 2017, Received in Revised Form: 31 March 2018, Accepted: 28 April 2018

Abstract

The use of genomic nucleotide sequences as biochemical signals in machine learning methods is possible by converting these sequences into numerical codes. This conversion results in an unrealistic increase in the dimension of the data and encounters some data analysis operations such as visualization and feature extraction with constraints. Therefore, one should use the dimensionality reduction technics in order to return the data to its real dimension. In this study, a deep autoencoder neural network has been used to reduce the dimension of binding site sequence data on the human genome. In order to determine whether the information of real data is preserved in compressed data, we perform a two-class classification using a support vector machine. The results show that information is almost entirely preserved in compression. Then, compressed data is used for visualization as well as feature selection by analysis of variance. The results show that the first, the tenth and eighth positions in the sequences are the most informative positions. While the majority of the previous works deal with gene expression data of microarrays and compare a few dimension reduction algorithms, this paper for the first time uses an autoencoder on nucleotide sequence data and provides a comprehensive comparison between the performance of the dimension reduction technics and machine learning algorithms.

Keywords: ***Autoencoder, Dimensionality Reduction, Genome Sequence, Classification, Feature Selection***

[*]**Corresponding Author**
Address: Bioelectric Department, Biomedical Engineering Faculty, Amirkabir University of Technology, P. O. Box: 158754413, Tehran, Iran
Tel: *+98-21-64542388*
Fax: *+98-21-64542388*
E-mail: *ssalehi@aut.ac.ir*

© Copyright 2018 ISBME, http://www.ijbme.org

مجله مهندسی پزشکی زیستی
دوره ۱۱، شماره ۳، پاییز ۱۳۹۶، ۲۳۰-۲۱۹
شناسه دیجیتال: 10.22041/ijbme.2018.75885.1294

کاهش بعد داده‌های توالی جایگاه‌های پیوند روی ژنوم انسان با استفاده از شبکه‌ی عصبی عمیق اتوانکودر

حسین بانکی کشکی[۱]، سیدعلی سیدصالحی[۲*]، فاطمه زارع میرک‌آباد[۳]

[۱] دانشجوی کارشناسی ارشد مهندسی پزشکی، گروه بیوالکتریک، دانشکده‌ی مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران
[۲] دانشیار، گروه بیوالکتریک، دانشکده‌ی مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران
[۳] استادیار، دانشکده‌ی ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران

تاریخ ثبت در سامانه: ۱۳۹۶/۸/۲۹، بازنگری: ۱۳۹۷/۱/۱۱، پذیرش قطعی: ۱۳۹۷/۲/۸

چکیده

استفاده از توالی‌های نوکلئوتیدی ژنوم به عنوان سیگنال‌های بیوشیمیایی در روش‌های یادگیری ماشین، با تبدیل این توالی‌ها به کدهای عددی امکان‌پذیر است و این تبدیل باعث افزایش غیرواقعی بعد داده‌ها شده و انجام عملیات‌های تحلیل داده، مانند بصری‌سازی و استخراج ویژگی را با محدودیت‌هایی روبه‌رو می‌سازد. از این‌رو، باید با استفاده از روش‌های کاهش بعد، داده‌ها را به فضای واقعی برگرداند. در این پژوهش از یک شبکه‌ی عصبی عمیق اتوانکودر به منظور کاهش بعد داده‌های توالی مربوط به جایگاه‌های پیوند روی ژنوم انسان استفاده شده است. به منظور بررسی میزان حفظ اطلاعات داده‌های اصلی در داده‌های کاهش بعد یافته، از یک طبقه‌بندی دوکلاسه به وسیله‌ی ماشین بردار پشتیبان استفاده می‌شود. نتایج به دست آمده نشان می‌دهد که اطلاعات تقریبا به طور کامل در فشرده‌سازی حفظ می‌شود. سپس از داده‌های فشرده‌شده برای بصری‌سازی و هم‌چنین انتخاب ویژگی با تحلیل واریانس استفاده می‌شود. نتایج به دست آمده نشان می‌دهد که مکان‌های اول، دهم و هشتم در توالی‌ها دارای بیشترین اطلاعات هستند. درحالی‌که عمده‌ی پژوهش‌های پیشین روی داده‌های بیان ژن حاصل از میکروآرایه، متمرکز شده‌اند و مقایسه‌ی محدودی بین روش‌های کاهش بعد در آن‌ها انجام شده است. این مقاله برای نخستین بار، داده‌های نوکلئوتیدی توالی را با شبکه‌ی اتوانکودر، کاهش بعد داده و مقایسه‌ی جامعی بین انواع روش‌های کاهش بعد و یادگیری ماشین ارائه می‌دهد.

کلیدواژه‌ها: اتوانکودر، کاهش بعد، توالی ژنوم، طبقه‌بندی، استخراج ویژگی

***نویسنده مسئول**
نشانی: گروه بیوالکتریک، دانشکده‌ی مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران، کد پستی: ۱۵۸۷۵۴۴۱۳
تلفن: ۶۴۵۴۲۳۸۸ (۲۱) ۹۸+
دورنگار: ۶۴۵۴۲۳۸۸ (۲۱) ۹۸+
پست الکترونیکی: ssalehi@aut.ac.ir

© Copyright 2018 ISBME, http://www.ijbme.org

۱- مقدمه

شناسایی جایگاه‌های پیوند عامل‌های رونویسی روی ژنوم انسان در سال‌های اخیر یکی از مسائل مهم در زمینه‌ی بیوانفورماتیک و روش‌های یادگیری ماشین از جمله شبکه‌های عصبی عمیق بوده است [3] [1]. عامل‌های رونویسی، پروتئین‌هایی هستند که آغاز یا خاتمه‌ی فرایند بیان ژن را کنترل می‌کنند. این پروتئین‌ها با قرارگیری روی محل‌های خاصی روی رشته‌ی دی.ان.ای، که جایگاه پیوند نامیده می‌شوند، عملیات فعال‌سازی یا غیرفعال‌سازی رونویسی را عهده‌دار هستند [4]. از آن‌جا که محل جایگاه‌های پیوند، اندکی قبل از محل ژن‌ها قرار گرفته است، کشف این جایگاه‌ها می‌تواند منجر به کشف ژن‌های جدید شود. تاکنون تعداد محدودی از جایگاه‌های پیوند با استفاده از روش‌های آزمایشگاهی شناسایی شده‌اند. هزینه‌بر بودن روش‌های آزمایشگاهی از یک‌سو و محدودیت‌های عملی از سوی دیگر، باعث شده است تا روش‌های یادگیری ماشین برای این منظور مورد استفاده قرار گیرند [5].

از آن‌جا که داده‌های جایگاه‌های پیوند به صورت توالی‌های نوکلئوتیدی (متشکل از نوکلئوتیدهای A و C و G و T) هستند، ابتدا باید به توالی‌های عددی تبدیل شوند تا پردازش روی آن‌ها توسط روش‌های یادگیری ماشین امکان‌پذیر گردد. کد کردن نوکلئوتیدها به صورت بیت‌های عددی به روش‌های مختلفی انجام می‌شود. در جدول (۱) نمونه‌هایی از این کدگذاری‌ها آورده شده است [6].

جدول(۱)- نمونه‌هایی از کدهای عددی برای توالی‌های نوکلئوتیدی

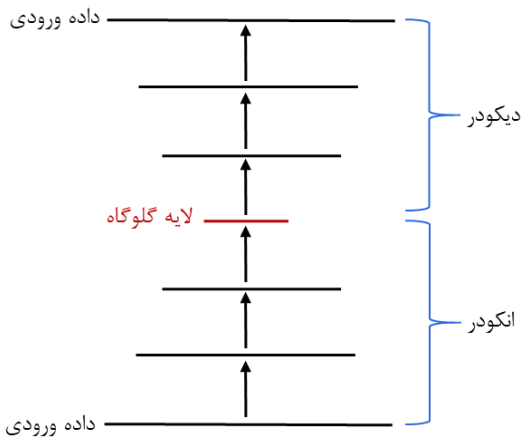
نوکلئوتید	کد عددی نوع۱	کد عددی نوع۲	کد عددی نوع۳
A	1 1 0 1	1 0 0 0	0 0
C	1 1 1 0	0 1 0 0	0 1
G	1 0 0 0	0 0 1 0	1 0
T	0 1 0 0	0 0 0 1	1 1

کد کردن توالی‌های نوکلئوتیدی به صورت بیت‌های عددی برای استفاده، هرچند گامی ضروری در روش‌های یادگیری ماشین است، اما محدودیت‌هایی را نیز در تحلیل داده‌ها ایجاد می‌کند، از جمله آن‌که بعد داده‌ها به صورت غیرواقعی افزایش پیدا می‌کند. در نتیجه، عملیات داده‌کاوی، مانند استخراج ویژگی و بصری‌سازی داده‌ها در فضای با بعد غیرواقعی امکان‌پذیر نمی‌شود. راه‌کار حل این مساله، کاهش بعد داده‌ها بدون از دست رفتن اطلاعات آن‌ها است، که برای این منظور می‌توان از شبکه‌های عصبی اتوانکودر استفاده نمود.

۱-۱- شبکه‌های عصبی اتوانکودر

اتوانکودرها نوعی از روش‌های کاهش بعد مبتنی بر شبکه‌های عصبی مصنوعی هستند که به دو گروه اتوانکودرهای خطی و غیرخطی تقسیم می‌شوند. اتوانکودرهای خطی فاقد نگاشت‌های غیرخطی هستند. اتوانکودرهای غیرخطی با دارا بودن نگاشت‌های غیرخطی به سه دسته‌ی بولین، شبکه‌های عصبی عمیق و ماشین‌های بولتزمان تقسیم می‌شوند. کاربرد اصلی اتوانکودرهای غیرخطی در کاهش بعد داده و حذف نویز می‌باشد [8] [7].

اتوانکودرها نخستین بار توسط کرامر، با عنوان شبکه‌های خودانجمنی، برای تحلیل مولفه‌های اساسی غیرخطی معرفی شدند [9]. این شبکه‌ها به دلیل ساختار ویژه‌ای که دارند، اغلب شبکه‌های گلوگاه نیز نامیده می‌شوند، که در آن‌ها داده‌های ورودی به تعداد متغیرهای کمتری در لایه‌ی گلوگاه فشرده می‌شوند. یک اتوانکودر از دو بخش اصلی انکودر و دیکودر تشکیل شده است. در قسمت انکودر، داده‌های ورودی در لایه‌ی گلوگاه، کاهش بعد پیدا می‌کنند و در قسمت دیکودر، داده‌ی کاهش بعد داده شده دوباره بازسازی می‌شود و به بعد اصلی باز می‌گردد (شکل ۱).



شکل(۱)- اجزای یک شبکه‌ی اتوانکودر

از آن‌جا که شبکه‌ی اتوانکودر چندلایه، دارای تعداد زیادی وزن است، تعلیم آن به روش پس‌انتشار خطا ممکن است منجر به هم‌گرایی شبکه نشود. از این‌رو بایستی از روش‌های پیش‌تعلیم برای بهینه‌سازی اولیه‌ی وزن‌ها استفاده کرد.

در این پژوهش، از اتوانکودرهای شبکه‌ی عمیق برای کاهش بعد داده‌های توالی جایگاه‌های پیوند روی ژنوم انسان استفاده شده است. در ابتدا، داده‌ها با اتوانکودر به بعد واقعی کاهش بعد داده شده‌اند، سپس با انجام طبقه‌بندی با استفاده از ماشین بردار

© Copyright 2018 ISBME, http://www.ijbme.org

پشتیبان[۱]، میزان حفظ اطلاعات در این داده‌ها مورد بررسی قرار گرفته است. در ادامه، عمل‌کرد اتوانکودر با سایر روش‌های کاهش بعد خطی و غیرخطی مقایسه شده و در نهایت، از داده‌های کاهش بعد یافته برای بصری‌سازی[۲] و استخراج ویژگی‌های[۳] مناسب استفاده شده است.

۱-۲- مرور پژوهش‌های پیشین

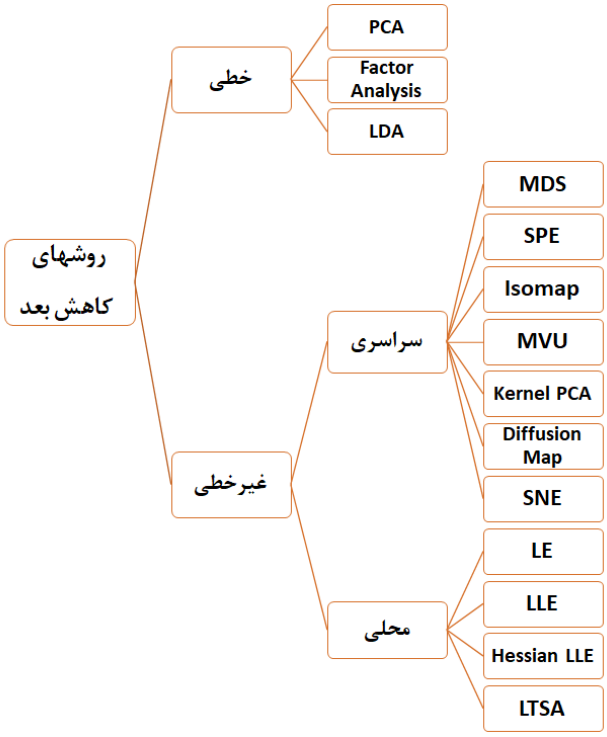
در این بخش، پژوهش‌های پیشین در دو قسمت مرور می‌شوند. در قسمت نخست، انواع روش‌های کاهش بعد (بدون توجه به زمینه‌ی کاربردی آن‌ها) شرح داده می‌شوند و در قسمت دوم، پژوهش‌های انجام شده با شبکه‌ی اتوانکودر در زمینه‌ی بیوانفوماتیک مورد بررسی قرار می‌گیرند.

۱-۲-۱- انواع روش‌های کاهش بعد

تاکنون روش‌های کاهش بعد مختلفی ارائه شده‌اند که در یک تقسیم‌بندی می‌توان آن‌ها را در دو گروه روش‌های خطی و غیرخطی دسته‌بندی کرد (شکل ۲). روش‌های غیرخطی خود به دو گروه روش‌های محلی[۴] و سراسری[۵] تقسیم می‌شوند. روش‌های سراسری، کاهش بعد را ضمن حفظ ویژگی مانیفلد همه‌ی داده‌ها انجام می‌دهند، درحالی‌که روش‌های محلی، ویژگی همسایگی‌های نزدیک هر داده را حفظ می‌کنند. تقسیم‌بندی دیگری نیز برای روش‌های کاهش بعد وجود دارد که بر اساس آن، این روش‌ها به دو دسته‌ی محدب و غیرمحدب تقسیم می‌شوند. در روش‌های کاهش بعد محدب، تابع هزینه‌ای بهینه می‌شود که دارای نقطه‌ی بهینه‌ی موضعی نیست، به عبارت دیگر، فضای جستجوی مساله، محدب است. به صورت برعکس، در روش‌های کاهش بعد غیرمحدب، تابع هزینه دارای نقطه‌ی بهینه‌ی موضعی است [11] [10].

روش‌های خطی، به‌عنوان روش‌های سنتی کاهش بعد، در مواجهه با داده‌های پیچیده کارایی مناسبی از خود نشان نمی‌دهند و دارای ضعف‌ها و مشکلاتی هستند. در کاهش بعد با روش تجزیه‌ی مولفه‌های اساسی[۶]، بایستی ماتریس کوواریانس برای داده‌های ورودی محاسبه شود. اندازه‌ی ماتریس کوواریانس به بعد داده‌های ورودی بستگی دارد. بنابراین، محاسبه‌ی بردارهای ویژه برای داده‌های با بعد بالا با این روش دشوار است. علاوه بر آن، تمرکز این روش بر فواصل دوتایی بزرگ در بین دادگان است و فواصل دوتایی کوچک در مانیفلد دادگان در نظر گرفته نمی‌شود [12]. تحلیل عاملی[۷]، بعد دادگان ورودی را کاهش داده و متغیرهای پنهان (فاکتور) را آشکار می‌سازد. متغیرهای پنهان لزوما در همه نوع داده‌ای وجود ندارند اما این روش برای هر نوع داده، حتی دادگان تصادفی، متغیرهای پنهان استخراج می‌کند [13]. کاهش بعد با تحلیل افتراقی خطی[۸] (LDA)، یک روش آماری بوده و نیازمند آن است که توزیع دادگان ورودی، نرمال باشد. از سوی دیگر، امکان تعیین دلخواه تعداد بعد برای دادگان در بعد پایین وجود ندارد [14].

روش‌های غیرخطی در مقایسه با روش‌های خطی، در مواجهه با دادگان واقعی که دارای مانیفلد غیرخطی و پیچیده هستند، کارایی بهتری دارند. هر یک از روش‌های غیرخطی کاهش بعد دارای مزایا و معایبی هستند که استفاده از آن‌ها را در کاربردهای مختلف با محدودیت‌هایی مواجه می‌سازد. به‌عنوان مثال، روش Isomap، MVU و LLE در صورت وجود ناپیوستگی در مانیفلد دادگان، نمی‌توانند عمل کاهش بعد را انجام دهند، یا در روش kernelPCA، اندازه‌ی ماتریس کرنل به مربع تعداد درایه‌ی ماتریس دادگان بستگی دارد و حجم محاسبات زیاد است [8].



شکل (۲)- انواع روش‌های خطی و غیرخطی کاهش بعد

[۱] Support Vector Machine (Svm)
[۲] Visualization
[۳] Feature Extraction
[۴] Local
[۵] Global
[۶] Principal Component Analysis (Pca)
[۷] Factor Analysis
[۸] Linear Discriminant Analysis (Lda)

© Copyright 2018 ISBME, http://www.ijbme.org

مشکلات روش‌های خطی و غیرخطی موجود باعث شده است که شبکه‌های عصبی اتوانکودر با قابلیت پردازش غیرخطی دادگان بعد بالا و پردازش مانیفلدهای غیرپیوسته، به عنوان ابزاری کارامد برای کاهش بعد شناخته شوند. در این پژوهش، از آن‌جا که دادگان توالی، واقعی هستند و احتمال غیرخطی بودن مانیفلد آن‌ها وجود دارد و از سوی دیگر، تبدیل عددی باعث شده است دادگان از فضای واقعی به فضای بعد بالاتر منتقل شوند (فضایی که در آن ممکن است مانیفلد دادگان غیرپیوسته باشد)، از شبکه‌ی اتوانکودر برای کاهش بعد استفاده شده است. شبکه‌های اتوانکودر هم‌چنین قادر هستند علاوه بر کاهش بعد دادگان، ویژگی‌های سطح بالا[1] را در لایه‌ی گلوگاه استخراج کنند و به نوعی عمل استخراج ویژگی را نیز انجام دهند. از سوی دیگر، اتوانکودرها قادر به حذف نویز از دادگان ورودی طی فرایند انکودینگ هستند و درصورتی‌که ثبت داده‌های آزمایشگاهی آغشته به نویز باشد، این نویز در لایه‌ی گلوگاه حذف می‌گردد.

۱-۲-۲- مرور پژوهش‌های پیشین با شبکه‌ی اتوانکودر

اخیرا پردازش داده‌های ژنوم با استفاده از شبکه‌های عمیق اتوانکودر در زمینه‌ی بیوانفوماتیک بسیار مورد توجه قرار گرفته است. چیکو و هم‌کاران در [15] از شبکه‌ی عمیق اتوانکودر برای کاهش بعد داده‌های مربوط به آنتالوژی ژن استفاده کرده، و سپس داده‌های کاهش بعد داده شده را برای پیش‌بینی عمل‌کرد ژن‌ها مورد استفاده قرار داده‌اند. در این پژوهش، عمل‌کرد شبکه‌ی اتوانکودر با روش tSVD (تجزیه به مقادیر منفرد) در کاهش بعد شش دسته‌ی داده، مورد مقایسه قرار گرفته و نشان داده شده است که عمل‌کرد اتوانکودر در مجموع بهتر است، زیرا اولا، اتوانکودر قابلیت تعلیم به صورت آنلاین روی مجموعه‌ی دادگان بزرگ را دارا است، ثانیا با استفاده از پردازنده‌های گرافیکی با سرعت بیشتری قابل تعلیم بوده، و ثالثا امکان کنترل پیچیدگی مدل با استفاده از تعداد و بزرگی لایه‌های پنهان فراهم است.

چن و هم‌کاران در [16]، از شبکه‌ی اتوانکودر برای یادگیری بازنمایی سلسله‌مراتبی فرایندهای رونویسی در مخمر استفاده کرده‌اند. در داخل سلول، فرایندهای دریافت و پردازش سیگنال‌های محیطی و درون‌سلولی به صورت سازمان‌یافته و سلسله‌مراتبی اتفاق می‌افتند، از این‌رو برای مدل کردن این فرایندها می‌توان از ساختار سلسله‌مراتبی اتوانکودر استفاده نمود. در این پژوهش، از داده‌های میکروآرایه‌ی مخمر به عنوان ورودی مدل اتوانکودر استفاده شده و تحلیل‌های زیستی با استفاده از مقادیر وزن‌ها در لایه‌های پنهان صورت گرفته است. نتایج به دست آمده نشان داده است که شبکه‌ی اتوانکودر به خوبی می‌تواند فرایند سیگنالینگ در سلول را مدل کند و از طرفی بازنمایی دقیق‌تر و سطح‌بالاتری را در اختیار قرار دهد.

تان و هم‌کاران در [17]، از شبکه‌ی اتوانکودر برای حذف نویز از داده‌های بیان ژن استفاده کرده‌اند. در این پژوهش، با اضافه کردن نویز تصادفی به داده‌های بیان ژن، نشان داده شده است که اتوانکودر می‌تواند نویز را حذف کرده و داده‌ها را مجددا بازسازی کند. مقایسه‌ی عمل‌کرد اتوانکودر با روش‌های کاهش بعد PCA و ICA نشان می‌دهد که حذف نویز به وسیله‌ی اتوانکودر به صورت بهتری انجام می‌گردد. تان و هم‌کاران در پژوهش دیگری در [18]، از شبکه‌ی اتوانکودر به عنوان ابزار یادگیری بدون سرپرستی برای استخراج ویژگی از داده‌های بیان ژن مربوط به سرطان سینه استفاده کرده و نشان داده‌اند که ویژگی‌های استخراج شده توسط اتوانکودر، با تفاسیر زیستی سازگاری دارد و می‌توان از این ویژگی‌ها برای تشخیص سرطان در سطح ژنوم استفاده کرد.

کوئی و هم‌کاران در [19]، از شبکه‌ی اتوانکودر برای کاهش بعد و استخراج ویژگی از داده‌های بیان ژن میکروآرایه استفاده کرده‌اند. در این پژوهش، ویژگی‌های استخراج شده و کاهش بعد داده شده برای خوشه‌بندی دادگان مورد استفاده قرار گرفته و نشان داده شده است که این ویژگی‌ها در مقایسه با ویژگی‌های اصلی، نتایج بهتری در خوشه‌بندی در اختیار می‌گذارند.

به صورت کلی، با توجه به پژوهش‌های انجام شده، کاربردهای شبکه‌های اتوانکودر را می‌توان در چهار دسته‌ی کاهش بعد، بازنمایی سلسله‌مراتبی، استخراج ویژگی و حذف نویز خلاصه کرد.

۲- مواد و روش‌ها

۲-۱- داده‌ها

داده‌های توالی ژنوم مورد استفاده (شامل ۲۳ جفت کروموزوم)، مربوط به نسخه‌ی ۱۸ام از ژنوم انسان[2] است که از پایگاه داده‌ی NCBI [20] استخراج شده است. تاکنون نسخه‌های جدیدتر مانند نسخه‌ی ۱۹ام[3] و نسخه‌ی ۳۸ام[4] نیز منتشر شده‌اند، اما از

[1] Abstract Features
[2] NCBI36/hg18, May 2006
[3] GRCh37/hg19, Feb 2009
[4] GRCh38/hg38, Dec 2013

© Copyright 2018 ISBME, http://www.ijbme.org

آن‌جا که حجم تحقیقات انجام شده پیرامون نسخه‌ی ۱۸ام گسترده‌تر از سایر نسخه‌ها است، در این پژوهش این نسخه مورد استفاده قرار گرفته است.

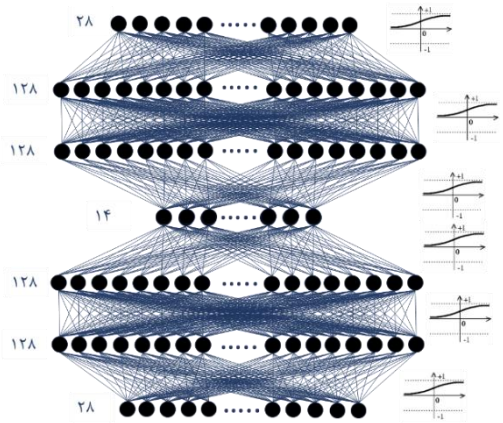
مکان جایگاه‌های پیوند عامل‌های رونویسی روی کروموزوم‌های مختلف از پایگاه داده‌ی فانتوم-۴ [21] استخراج شده و سپس با استفاده از زبان برنامه‌نویسی Perl، توالی‌های جایگاه‌های پیوند مربوط به هر عامل به دست آمده است. تعداد عامل‌های رونویسی شناخته شده روی نسخه‌ی ۱۸ام ژنوم انسان برابر ۲۰۰ عامل و مجموع جایگاه‌های پیوند این عامل‌ها برابر ۲۲۸۰۱۸ جایگاه است. بیشترین تعداد جایگاه‌های پیوند به ترتیب مربوط به عامل رونویسی SP1 با ۱۷۹۶۵ جایگاه، عامل MAZ با ۹۱۰۸ جایگاه و عامل ELF1,2,4 با ۵۳۰۳ جایگاه است. طول جایگاه‌های پیوند از ۴ تا ۳۰ نوکلئوتید متغیر است.

در این مقاله، توالی جایگاه‌های پیوند عامل SP1 روی کروموزوم-۱ که شامل ۱۹۰۳ توالی است، مورد استفاده قرار می‌گیرد. به تعداد جایگاه‌های پیوند، جایگاه‌های غیرپیوند نیز از روی کروموزوم-۱ برداشته می‌شوند، بنابراین مجموع دادگان برابر ۳۸۰۶ توالی و شامل دو کلاس داده است (جایگاه‌های پیوند، جایگاه‌های غیرپیوند). توالی‌های نوکلئوتیدی با استفاده از کدگذاری نوع-۳ در جدول (۱) به کد عددی تبدیل می‌شوند، بنابراین بعد داده‌ها از ۱۴ به ۲۸ افزایش می‌یابد (جدول ۲).

۲-۲- تعلیم شبکه‌ی عمیق اتوانکودر

شبکه‌ی عمیق اتوانکودر مورد استفاده از پنج لایه‌ی پنهان تشکیل شده است. لایه‌ی ورودی و خروجی دارای ۲۸ نورون، لایه‌ی پنهان اول، دوم، چهارم و پنجم دارای ۱۲۸ نورون و لایه‌ی گلوگاه دارای ۱۴ نورون است (شکل ۳). تابع انتقال از نوع تک‌قطبی (logsig) و الگوریتم تعلیم پس‌انتشار خطا است.

در شبکه‌های عصبی عمیق، مشکل همگرا نشدن شبکه (عدم کاهش خطای تعلیم) به علت کاهش گرادیان در لایه‌های پایین‌تر، یک مشکل عمده است. از سوی دیگر با عمیق‌تر شدن شبکه، وابستگی زیادی مابین پارامترهای آن در حین تعلیم به وجود می‌آید و نقش کمینه‌های موضعی پررنگ می‌شود، چراکه تعداد کمینه‌های موضعی فعال که به ساختار شبکه و مقداردهی اولیه‌ی پارامترهای آن بستگی دارند، بیشتر است [23] [22]. برای جلوگیری از این مشکلات، از روش‌های پیش‌تعلیم استفاده می‌شود. روش‌های پیش‌تعلیم، باعث انتقال وزن‌های شبکه به نقطه‌ی بهینه می‌شوند و شبکه هنگام تعلیم به جای شروع از وزن‌های اولیه تصادفی، از وزن‌های بهینه شروع به تعلیم می‌کند. روش‌های پیش‌تعلیم مختلفی برای شبکه‌های عصبی عمیق معرفی شده‌اند که از جمله‌ی آن‌ها می‌توان به الگوریتم ژنتیک، روش مشرف به هدف و تجزیه‌ی شبکه به تعدادی ماشین بولتزمان اشاره کرد. در این پژوهش، از روش پیش‌تعلیم معرفی شده درمنبع [24] استفاده شده است که به روش پیش‌تعلیم لایه‌به‌لایه معروف است.



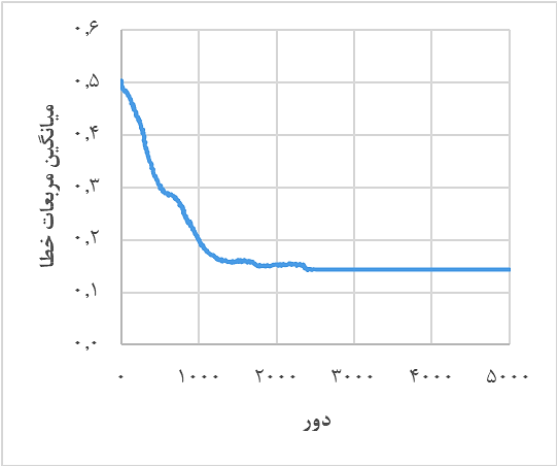
شکل (۳)- ساختار شبکه‌ی اتوانکودر مورد استفاده

جدول (۲)- توالی نوکلئوتیدی جایگاه‌های پیوند و غیرپیوند و کد عددی معادل آن‌ها، برچسب اختصاص یافته به هر کلاس

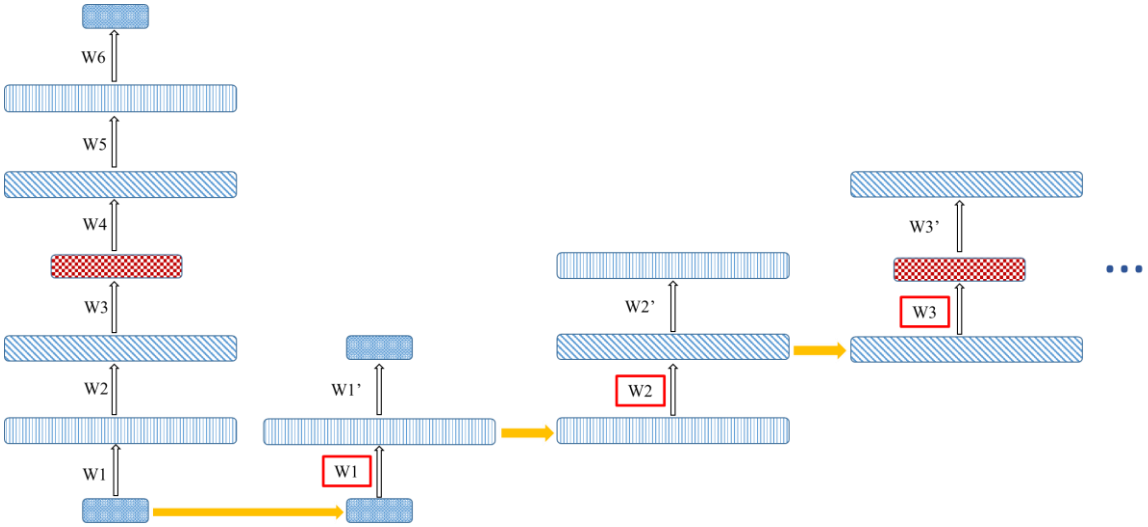
	توالی نوکلئوتیدی	کد عددی	برچسب
جایگاه‌های پیوند	AGGGGCGGGGCGCG	00101010100110101010011001 10	0
	GTGGGTGGGGGCGC	10111010101110101010100110 01	0
	GGGGGAGGGGCGCC	10101010100010101010011001 00	0
	GGCGGGGCGGGGGG	10100110101010011010101010	0
	...	...	...
جایگاه‌های غیرپیوند	CTATTCAAATAACT	01110011110100000011000001 11	1
	GGATCTGCTATTCA	10100011011110011100111101 00	1
	AAATGAATTGAGGT	00000011100000111110001010 11	1
	GATGTATCTCTTAA	10001110110011011101111100 00	1

© Copyright 2018 ISBME, http://www.ijbme.org

پیش‌تعلیم لایه‌به‌لایه در واقع شکستن مساله‌ی پیچیده‌ی یادگیری عمیق به چند زیرمساله‌ی یادگیری سطحی است (شکل ۴). در این روش، شبکه‌ی عمیق به چند شبکه‌ی ساده‌ی تک‌لایه‌ی پنهان شکسته می‌شود و هر کدام از شبکه‌های ساده به تنهایی با الگوریتم پس‌انتشار خطا تعلیم دیده و همگرا می‌شوند. در نهایت، وزن‌های لایه‌ی اول شبکه‌های ساده، وزن‌های اولیه‌ی شبکه‌ی عمیق را تشکیل می‌دهند. پس از پیش‌تعلیم شبکه و به دست آمدن وزن‌های اولیه‌ی بهینه به روش فوق، شبکه با روش پس‌انتشارخطا تعلیم داده می‌شود. نمودار خطای شبکه در شکل (۵) نشان داده شده است. از دور ۲۲۰۰ به بعد، شبکه همگرا شده و خطا ثابت می‌ماند.



شکل (۵)- نمودار خطای تعلیم شبکه



شکل (۴)- شماتیک نحوه‌ی پیش‌تعلیم لایه‌به‌لایه‌ی شبکه‌ی اتوانکودر

۳- یافته‌ها و بحث

۳-۱- صحت‌سنجی دادگان کاهش بعد یافته

نمودار خطای شبکه به عنوان معیاری برای سنجش تعلیم صحیح آن به کار می‌رود، اما برای بررسی حفظ اطلاعات دادگان اصلی در دادگان فشرده شده نمی‌توان از این معیار استفاده کرد. در این پژوهش، برای سنجش حفظ اطلاعات دادگان اصلی در دادگان فشرده شده، از روش طبقه‌بندی ماشین بردار پشتیبان استفاده می‌شود. با استفاده از این روش، طبقه‌بندی دو کلاسه روی جایگاه‌های پیوند و جایگاه‌های غیرپیوند، یک بار با دادگان اصلی و بار دیگر با دادگان فشرده شده انجام می‌شود. درصورتی‌که درصد صحت طبقه‌بندی در دادگان اصلی برابر یا نزدیک به درصد صحت طبقه‌بندی در دادگان فشرده شده باشد، می‌توان نتیجه گرفت که اطلاعات در دادگان فشرده شده حفظ شده است.

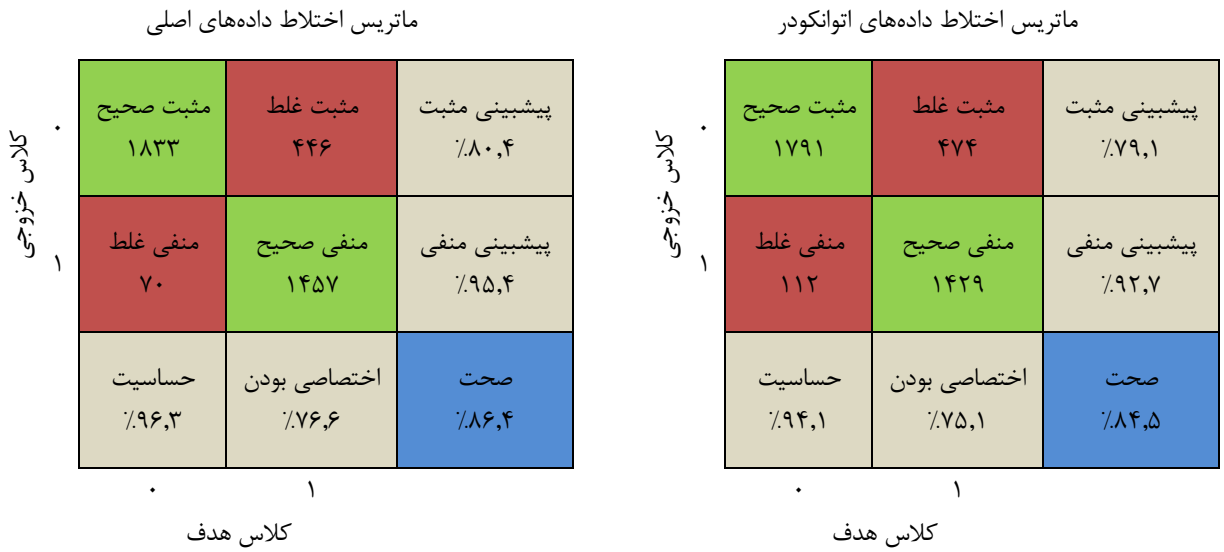
ماتریس‌های اختلاط نشان داده شده در شکل (۶)، درصد صحت به دست آمده از طبقه‌بندی را روی دو دسته‌ی داده‌ی مذکور نشان می‌دهد. بر این اساس، درصد صحت روی دادگان اصلی برابر ۸۶,۴٪ و روی دادگان فشرده‌شده برابر ۸۴,۵٪ است. نزدیکی درصد صحت دو دسته‌ی داده به معنی حفظ اطلاعات در دادگان فشرده است. علاوه بر درصد صحت، نتایج دیگر مانند حساسیت، اختصاصی بودن، دقت و مقدار پیش‌بینی منفی و مثبت نیز به یک‌دیگر نزدیک هستند.

۳-۲- مقایسه‌ی اتوانکودر با سایر روش‌های کاهش بعد

علاوه بر شبکه‌ی عمیق اتوانکودر، با استفاده از جعبه‌ابزار کاهش بعد در نرم‌افزار متلب، سایر روش‌های کاهش بعد روی دادگان اصلی اعمال می‌شود. در این روش‌ها نیز بعد اصلی ۱۴ در نظر گرفته شده است. سپس روی داده‌های کاهش بعد یافته، با

© Copyright 2018 ISBME, http://www.ijbme.org

الگوریتم‌های یادگیری ماشین (با استفاده از جعبه‌ابزار یادگیری ماشین متلب)، طبقه‌بندی انجام می‌شود. نتایج درصد صحت در جدول (۳) نشان داده شده است. بر این اساس، در بین روش‌های کاهش بعد، روش‌های Kernel PCA، Diffusion Map و Factor Analysis، در مجموع نسبت به سایر روش‌ها درصد صحت بیشتری در اختیار می‌گذارند. با این وجود، بیشترین درصد صحت مربوط به روش PCA با مقدار ۸۶٫۲٪ است. در بین روش‌های یادگیری ماشین، روش ماشین بردار پشتیبان در حالت گاوسین و کوادراتیک، بهترین عمل‌کرد را در طبقه‌بندی داشته‌اند.



شکل (۶)- ماتریس اختلاط مربوط به طبقه‌بندی داده‌های اصلی و داده‌های کاهش بعد یافته‌ی اتوانکودر

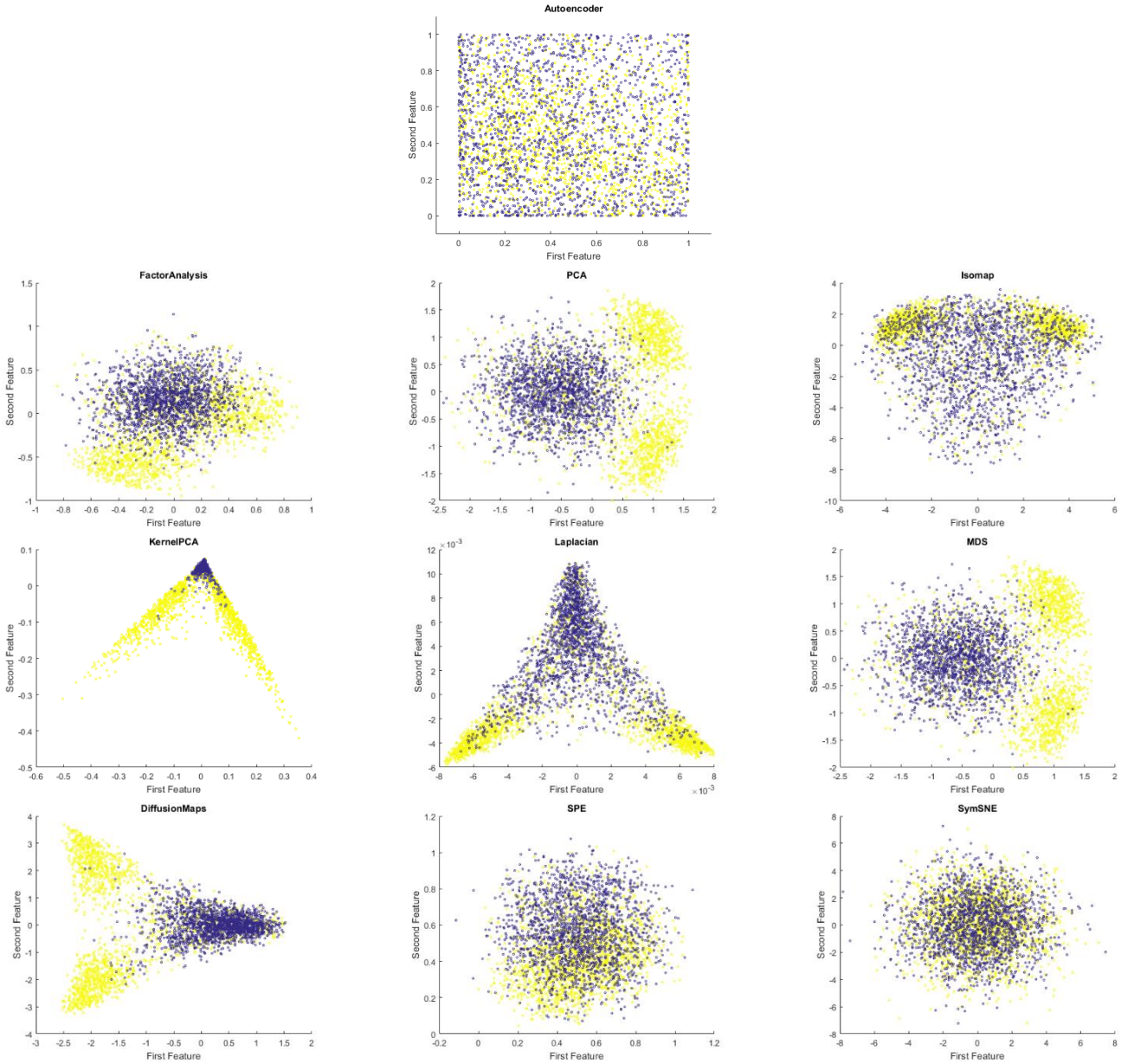
جدول (۳)- درصد صحت طبقه‌بندی داده‌های کاهش بعد یافته با استفاده از روش‌های یادگیری ماشین

	Real Data	Autoencoder	MDS	SPE	Isomap	Kernel PCA	Diffusoin Map	Laplacian	Sym SNE	PCA	Factor Analysis
C_Tree	۸۰٫۸	۷۷.۷	۸۱.۹	۷۸.۲	۷۲.۶	۸۲.۳	۸۳.۳	۷۶.۹	۴۹.۶	۸۲	۸۱.۵
M_Tree	۸۲	۷۹	۸۳.۹	۷۸.۴	۷۲.۲	۸۴.۶	۸۴.۴	۷۹.۱	۴۹.۴	۸۳.۷	۸۲.۸
S_Tree	۸۰.۱	۷۵.۵	۸۴.۲	۷۴.۹	۷۱.۵	۸۵.۲	۸۴.۷	۸۱.۱	۴۹.۶	۸۴	۸۲.۳
LDA	۸۱.۸	۷۹.۶	۸۱.۹	۸۱	۶۹.۶	۸۴.۱	۸۴.۴	۷۷.۸	۵۰.۸	۸۲	۱.۸۱
QDA	۸۳.۶	۸۰.۲	۸۲.۹	۸۲.۴	۷۴.۶	۸۵.۵	۸۲.۲	۷۹.۹	۴۹.۹	۸۳	۸۲.۶
RL	۸۲	۸۰.۱	۸۱.۸	۸۱.۳	۶۹.۷	۸۵.۱	۸۳.۷	۷۸.۲	۵۰.۳	۸۱.۷	۸۱.۸
L_SVM	۸۳.۴	۸۱.۷	۸۳.۶	۸۳.۱	۶۸.۸	۸۵.۱	۸۴.۶	۷۸.۸	۵۰.۸	۸۳.۷	۸۳.۶
Q_SVM	۸۵.۷	۸۳.۶	۸۶	۸۴.۸	۷۵.۴	۸۴.۹	۸۵.۱	۸۱.۵	۵۰	۸۵.۸	۸۵.۲
C_SVM	۸۳.۸	۸۱.۱	۸۲.۸	۸۰.۸	۷۲.۸	۸۵.۱	۸۱.۲	۸۰.۶	۴۸.۴	۸۲.۱	۸۴.۴
FG_SVM	۶۹	۷۲.۴	۷۱.۹	۷۳.۵	۷۶.۶	۸۵.۹	۸۴.۳	۷۸.۱	۵۰.۱	۷۱.۸	۸۲.۵
MG_SVM	۸۶.۴	۸۴.۵	۸۵.۹	۸۵.۴	۷۷.۷	۸۵.۳	۸۵.۵	۸۱.۱	۴۸.۲	۸۶.۲	۸۵.۷
CG_SVM	۸۴.۲	۸۲.۳	۸۴	۸۳.۷	۷۰.۵	۸۴.۷	۸۴.۸	۸۰	۵۰.۱	۸۴.۱	۸۵.۱
F_KNN	۸۰.۱	۷۹.۹	۸۰.۹	۷۹.۲	۶۹	۸۰.۴	۷۹.۸	۷۵.۴	۵۰.۶	۸۰.۱	۸۱.۱
M_KNN	۸۲.۶	۸۲.۶	۸۴.۵	۸۳.۳	۷۱.۶	۸۳.۴	۸۴	۸۰.۳	۴۹.۸	۸۳.۹	۸۵.۵
Co_KNN	۸۳.۹	۸۰.۱	۸۴.۳	۸۴.۳	۷۴.۷	۸۵	۸۲.۵	۷۹.۸	۵۰	۸۳.۵	۸۵.۳
Cu_KNN	۸۲.۶	۸۳.۳	۸۴.۲	۸۲.۹	۷۱.۶	۸۴	۸۴.۸	۷۹.۷	۴۷.۸	۸۳.۸	۸۵.۳
W_KNN	۸۲.۴	۸۲.۵	۸۴.۵	۸۳	۷۱.۵	۸۴.۵	۸۴	۸۰.۷	۷.۴۹	۸۴.۱	۸۵.۵

© Copyright 2018 ISBME, http://www.ijbme.org

۳-۳- بصری‌سازی داده‌ها

کاهش بعد داده‌ها به بعد واقعی، امکان بصری‌سازی داده‌ها با استفاده از ابعاد را فراهم می‌سازد. بصری‌سازی ابعاد (ویژگی‌ها) ضمن ارائه‌ی درک از ماهیت داده‌ها، می‌تواند در انتخاب ویژگی نیز مورد استفاده قرار گیرد. می‌توان با رسم ویژگی‌ها بر حسب یک‌دیگر، قدرت ایجاد تفکیک‌پذیری بین کلاس‌ها توسط ویژگی‌ها را نمایش داد. در شکل (۷)، تفکیک‌پذیری دو کلاس (جایگاه‌های پیوند و جایگاه‌های غیرپیوند) بر اساس ویژگی‌های اول و دومِ به دست آمده از اتوانکودر و سایر روش‌های کاهش بعد نشان داده شده است. این نمودارها برای سایر ویژگی‌ها و حتی به صورت سه‌بعدی نیز قابل نمایش هستند. با مقایسه‌ی نمودارهای تفکیک‌پذیری روش‌های کاهش بعد با نتایج به دست آمده در جدول (۳)، می‌توان نتیجه گرفت که آن دسته از روش‌های کاهش بعد که تفکیک‌پذیری بیشتری بین داده‌ها ایجاد می‌کنند، درصد صحت بالاتری در طبقه‌بندی در اختیار می‌گذارند. در نمودارهای شکل (۷)، نقاط زردرنگ مربوط به جایگاه‌های پیوند و نقاط آبی رنگ مربوط به جایگاه‌های غیرپیوند هستند. در اغلب روش‌ها، نقاط زرد رنگ (جایگاه‌های پیوند) به دو مجموعه‌ی جدا از هم به صورت منظم تفکیک شده‌اند، اما نقاط آبی رنگ (جایگاه‌های غیرپیوند) به صورت پراکنده در نمودار قرار دارند. دلیل این پدیده این است که جایگاه‌های غیرپیوند به صورت تصادفی از روی کروموزوم برداشته شده‌اند اما جایگاه‌های پیوند، جایگاه‌های واقعی هستند که با روش‌های آزمایشگاهی استخراج شده‌اند.



شکل (۷)- نمودار میزان تفکیک‌پذیری کلاس‌ها با ویژگی اول و دوم در روش‌های کاهش بعد مختلف

© Copyright 2018 ISBME, http://www.ijbme.org

۳-۴- انتخاب ویژگی

هرچند انتخاب ویژگی مناسب توسط نمودارهای تفکیک‌پذیری ویژگی‌ها تا حدودی امکان‌پذیر است، اما برای اطمینان بایستی از روش‌های انتخاب ویژگی مانند تحلیل واریانس استفاده کرد. تحلیل واریانس یک روش آماری است که به وسیله‌ی آن می‌توان وجود تفاوت معنادار آماری بین میانگین متغیرهای توصیف‌کننده‌ی دو یا چندین گروه (کلاس) را بررسی کرد. این آزمون، درجه‌ی اهمیت هر یک از متغیرها (ویژگی‌ها) را در ایجاد تمایز بین گروه‌ها با مقدار p-value گزارش می‌کند. متغیرهای با مقدار p-value کمتر، بیشترین تفکیک‌پذیری را بین گروه‌ها ایجاد می‌کنند. انجام تحلیل واریانس روی داده‌های فشرده-شده‌ی لایه‌ی گلوگاه اتوانکودر برای به دست آوردن مکان‌های پراهمیت، نشان می‌دهد که مکان‌های اول، دهم، هشتم و پنجم به ترتیب پراهمیت‌ترین مکان‌ها می‌باشند. جدول (۴) درجه‌ی اهمیت هر مکان را با مقدار p-value نشان می‌دهد.

ذکر این نکته ضروری است که در تحلیل آماری با استفاده از مقدار p-value، معمولا، مقایسه‌ی مقدار p-value با یک مقدار مشخص برای اثبات یا رد یک فرض (فرض صفر) رایج است. به عنوان مثال، در اغلب پژوهش‌های آماری، در صورتی که مقدار p-value بیش از ۰٫۰۵ باشد، فرض صفر رد می‌شود و در صورتی‌که کمتر از ۰٫۰۵ باشد، فرض صفر ثابت می‌گردد. در این پژوهش، از مقدار p-value برای استخراج ویژگی‌های پراهمیت استفاده می‌شود. به عبارت دیگر، فرض صفر برای هر ویژگی، این است که آن ویژگی دارای اهمیت آماری است. با توجه به جدول (۴)، تمامی مقادیر کمتر از ۰٫۰۵ هستند، پس معنی‌دار و دارای اهمیت می‌باشند. به صورت رایج، مقدار p-value تا سه رقم اعشار گزارش می‌گردد. در جدول (۴)، مقادیر با تعداد اعداد اعشار بیشتری گزارش شده‌اند تا تفاوت آن‌ها کاملا مشخص شود.

با انجام طبقه‌بندی، یک‌بار با استفاده از ویژگی‌های با اهمیت (مکان‌های اول، دهم، هشتم، پنجم، نهم، سیزدهم و هفتم)، و بار دیگر با ویژگی‌های کم‌اهمیت (مکان‌های چهارم، سوم، دوم، یازدهم، چهاردهم، دوازدهم و ششم)، مشاهده می‌شود که درصد صحت بین این دو حالت حدود ۱۰٪ تفاوت دارد و ویژگی‌های پراهمیت درصد صحت بیشتری نتیجه می‌دهند. شکل (۸) درصد صحت طبقه‌بندی با هر دسته از ویژگی‌ها را با استفاده از طبقه‌بندی SVM نشان می‌دهد.

جدول (۴)- اهمیت ویژگی‌ها بر حسب مقدار p-value به دست آمده از تحلیل واریانس

مکان نوکلئوتیدی	مقدار p-value
اول	۰.۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰
دهم	۰.۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰
هشتم	۰.۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰۰
پنجم	۰.۰۰۰۰۰۰۰۰۰۰۰۰۰۱۰۸۶
نهم	۰.۰۰۰۰۰۰۳۳۴۴۹۰۱۹۰۰
سیزدهم	۰.۰۰۰۰۰۰۴۱۴۰۴۸۷۳۹۸
هفتم	۰.۰۰۰۰۰۰۸۵۴۹۴۱۰۸۸۸
چهارم	۰.۰۰۰۰۰۱۹۸۲۶۶۵۷۱۲۴
سوم	۰.۰۰۰۰۰۴۰۸۷۳۷۹۳۹۰۵
دوم	۰.۰۰۰۰۶۰۷۰۳۱۱۱۸۹۱۶
یازدهم	۰.۰۰۰۰۶۷۰۴۲۳۱۵۶۵۲۰
چهاردهم	۰.۰۰۱۴۹۹۴۴۵۹۱۳۰۱۲۴
دوازدهم	۰.۰۲۳۰۶۹۱۸۰۵۴۹۵۶۶۶
ششم	۰.۰۳۱۶۲۵۹۲۷۹۳۲۹۳۳۲

شکل (۸)- نمودار درصد صحت طبقه‌بندی با استفاده از ویژگی‌های پراهمیت و کم‌اهمیت

۴- نتیجه‌گیری

در این پژوهش، مساله‌ی افزایش بعد دادگان توالی ژنوم انسان و نحوه‌ی بازگرداندن داده‌ها به بعد واقعی مورد بررسی قرار گرفت. روش‌های مختلف کد کردن عددی توالی‌ها برای استفاده در الگوریتم‌های یادگیری ماشین باعث افزایش بعد غیرواقعی دادگان می‌شود و تحلیل دادگان در فضای غیرواقعی محدودیت‌هایی مانند عدم توانایی بصری‌سازی و انتخاب ویژگی را ایجاد می‌کند.

© Copyright 2018 ISBME, http://www.ijbme.org

پژوهش‌های پیشین با شبکه‌ی عمیق اتوانکودر برای کاهش بعد داده‌های ژنوم، عمدتا روی داده‌های بیان ژن حاصل از میکروآرایه بوده‌اند، که این داده‌ها خود به صورت عددی و در بعد واقعی هستند. در این پژوهش، داده‌ها به صورت توالی نوکلئوتیدی بوده و عملیات کاهش بعد نیز روی داده‌های تبدیل یافته در یک بعد غیرواقعی انجام شده است. از سوی دیگر در پژوهش‌های پیشین، مقایسه‌ی عمل‌کرد شبکه‌ی اتوانکودر عمدتا با روش PCA یا تعداد محدودی روش دیگر انجام شده است، اما در این پژوهش سعی شده است مقایسه‌ی جامعی بین عمل‌کرد شبکه‌ی اتوانکودر و اغلب روش‌های خطی و غیرخطی ارائه گردد. علاوه بر آن، مقایسه‌ی جامعی بین روش‌های یادگیری ماشین ارائه شده است. به صورت خلاصه، نوآوری این پژوهش را می‌توان از دو منظر "نوع داده‌ها" و "مقایسه‌ی انواع روش‌های کاهش بعد" دانست.

در این پژوهش، از یک شبکه‌ی عمیق اتوانکودر پنج لایه برای کاهش بعد دادگان از فضای ۲۸ بعدی به فضای ۱۴ بعدی استفاده شد. برای جلوگیری از عدم همگرایی شبکه در حین تعلیم، به جای استفاده از وزن‌های اولیه‌ی تصادفی، از وزن‌های به دست آمده از روش پیش‌تعلیم استفاده شد.

به منظور سنجش میزان اطلاعات در داده‌های فشرده شده، از طبقه‌بندی دوکلاسه‌ی ماشین بردار پشتیبان استفاده شد، و نتایج به دست آمده نشان داد که اطلاعات تا حد زیادی در لایه‌ی گلوگاه حفظ می‌شوند و تفاوت درصد صحت طبقه‌بندی در داده‌های فشرده شده با داده‌های اصلی فقط ۱٫۹٪ درصد است. مقایسه‌ی عمل‌کرد شبکه‌ی اتوانکودر با سایر روش‌های کاهش بعد نشان داد که بیشینه‌ی درصد صحت به دست آمده با اتوانکودر با بیشینه‌ی درصد به دست آمده در سایر روش‌های کاهش بعد (که مربوط به روش PCA است) فقط ۱٫۷٪ اختلاف دارد. انتخاب ویژگی با تحلیل واریانس روی دادگان فشرده شده‌ی اتوانکودر نشان داد که مکان‌های اول، دهم، هشتم، پنجم، نهم، سیزدهم و هفتم از توالی نوکلئوتیدی به ترتیب به عنوان مکان‌های پراهمیت برای عامل‌های رونویسی محسوب می‌شوند و درصد صحت طبقه‌بندی با این مکان‌ها (ویژگی‌ها) حدود ۱۰٪ بیشتر از درصد صحت طبقه‌بندی با ویژگی‌های کم اهمیت است.

هرچند عمل‌کرد اتوانکودر در کاهش بعد داده‌ها در این پژوهش با عمل‌کرد بهترین روش‌های کاهش بعد تفاوت اندکی داشت (با توجه به معیار درصد صحت طبقه‌بندی)، اما قدرت تفکیک‌پذیری بین طبقه‌ها به هنگام بصری‌سازی کمتر بود. برای بهبود قدرت تفکیک‌دهندگی اتوانکودر در کاهش بعد داده‌ها می‌توان از روش‌هایی مانند کاهش بعد با سرپرستی استفاده کرد و کارایی آن‌ها را بررسی نمود. از آن‌جاکه در این روش‌ها برچسب داده‌ها نیز در فرایند کاهش بعد دخیل می‌شود، امکان افزایش قدرت تفکیک‌دهندگی وجود دارد.

۵- سپاس‌گزاری

بدین وسیله از زحمات جناب آقای دکتر ابوذر غفاری، مدرس درس داده‌کاوی پزشکی در دانشکده‌ی مهندسی پزشکی، که با زحمات بی‌دریغ خود این‌جانب را در کسب مهارت در علم تحلیل داده یاری نمودند، صمیمانه سپاس‌گزاری می‌گردد.

۶- مراجع

[1] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, “Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning,” *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015.

[2] D. R. Kelley, J. Snoek, and J. L. Rinn, “Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks,” *Genome Res.*, vol. 26, no. 7, pp. 990–999, 2016.

[3] S. Zhang *et al.*, “A deep learning framework for modeling structural features of RNA-binding protein targets,” *Nucleic Acids Res.*, vol. 44, no. 4, pp. 1–14, 2015.

[4] S. Inukai, K. H. Kock, and M. L. Bulyk, “Transcription factor–DNA binding: beyond binding site motifs,” *Curr. Opin. Genet. Dev.*, vol. 43, pp. 110–119, 2017.

[5] N. Jayaram, D. Usvyat, and A. C. R. Martin, “Evaluating tools for transcription factor binding site prediction,” *BMC Bioinformatics*, no. i, pp. 1–12, 2016.

[6] D. Liu, X. Xiong, B. DasGupta, and H. Zhang, “Motif discoveries in unaligned molecular sequences using self-organizing neural networks,” *IEEE Trans. Neural Networks*, vol. 17, no. 4, pp. 919–928, 2006.

[7] G. E. Hinton, “Reducing the Dimensionality of Data with Neural Networks,” *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, 2006.

[8] Y. Bengio, “Learning Deep Architectures for AI,” *Found. Trends® Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[9] M. A. Kramer, “Nonlinear principal component analysis using autoassociative neural networks,” *AIChE J.*, vol. 37, no. 2, pp. 233–243, 1991.

[10] L. J. P. Van der Maaten and U. Maastricht, “An introduction to dimensionality reduction using matlab,” *Report*, vol. 1201, no. 07–07, p. 62, 2007.

[11] L. J. P. Van Der Maaten, E. O. Postma, and H. J. Van Den Herik, “Dimensionality Reduction: A

© Copyright 2018 ISBME, http://www.ijbme.org

[18] J. Tan, M. Ung, C. Cheng, and C. S. Greene, "Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders," in *Pacific Symposium on Biocomputing Co-Chairs*, 2014, pp. 132–143.

[19] H. Cui, C. Zhou, X. Dai, Y. Liang, R. Paffenroth, and D. Korkin, "Boosting Gene Expression Clustering with System-Wide Biological Information: A Robust Autoencoder Approach," *bioRxiv*, p. 214122, 2017.

[20] "NCBI Genome Database," 2017. [Online]. Available: http://hgdownload.cse.ucsc.edu/goldenPath/hg18/chromosomes/. [Accessed: 11-Mar-2017].

[21] "FANTOM4 TFBS Data," 2017. [Online]. Available: http://fantom.gsc.riken.jp/4/download/GenomeBrowser/hg18/TFBS_CAGE/.

[22] Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7700 LECTU, pp. 437–478, 2012.

[23] D. Erhan, A. Courville, and P. Vincent, "Why Does Unsupervised Pre-training Help Deep Learning ?," *J. Mach. Learn. Res.*, vol. 11, pp. 625–660, 2010.

[24] S. Z. Seyyedsalehi and S. A. Seyyedsalehi, "A fast and efficient pre-training method based on layer-by-layer maximum discrimination for deep neural networks," *Neurocomputing*, vol. 168, pp. 669–680, 2015.

Comparative Review," *J. Mach. Learn. Res.*, vol. 10, pp. 1–41, 2009.

[12] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[13] I. Pournara and L. Wernisch, "Factor analysis for gene regulatory networks and transcription factor activity profiles," *BMC Bioinformatics*, vol. 8, no. 1, p. 61, 2007.

[14] A. Sharma and K. K. Paliwal, "Cancer classification by gradient LDA technique using microarray gene expression data," *Data Knowl. Eng.*, vol. 66, no. 2, pp. 338–347, 2008.

[15] D. Chicco, P. Sadowski, and P. Baldi, "Deep autoencoder neural networks for gene ontology annotation predictions," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2014, pp. 533–540.

[16] L. Chen, C. Cai, V. Chen, and X. Lu, "Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model," in *BMC bioinformatics*, 2016, vol. 17, no. 1, p. S9.

[17] J. Tan, J. H. Hammond, D. A. Hogan, and C. S. Greene, "ADAGE-based integration of publicly available Pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions," *MSystems*, vol. 1, no. 1, pp. e00025-15, 2016.

© Copyright 2018 ISBME, http://www.ijbme.org