دومین کنفرانس ملی تحقیقات کاربردی در مهندسی برق کامپیوتر فناوری اطلاعات
شیراز ۱۹ مرداد ماه ۱۳۹۶
سالن همایش های فرهنگیان
The Second Annual Conference on Applied Research in Electrical, Computer and IT Engineering
2017

ارزیابی تاثیر کاهش داده در عملکرد طبقه بندی کننده های خطی و غیرخطی

حمید محمودآبادی[1]، دکتر محمدرضا اصغری اسکویی[2]

[1] دانشگاه علامه طباطبایی، Mahmoodabadihamid@atu.ac.ir

[2] دانشگاه علامه طباطبایی، Oskoei@atu.ac.ir

چکیده:

بررسی و مقایسه ی عملکرد الگوریتم های مختلف طبقه بندی در مسائل کاربردی یادگیری ماشین و شناخت الگو، همواره یک چالش بوده است. همچنین امروزه با افزایش میزان داده های تولید شده، در به کار گیری الگوریتم های طبقه بندی کننده، با مشکل بالا بودن حجم محاسبات مواجه هستیم. به همین علت، بکارگیری الگوریتم هایی که در پی کاهش ابعاد داده ها هستند، مورد توجه بسیاری قرار گرفته است. ما در این مقاله قصد داریم میزان پایداری و عملکرد چهار الگوریتم طبقه بندی کننده بیزین، نزدیکترین همسایه، ماشین بردار پشتیبان و پرسپترون چند لایه را بر روی مجموعه داده IRIS، در برابر کاهش ابعاد داده ها توسط الگوریتم PCA بررسی کنیم.

واژه‌های کلیدی

ارزیابی طبقه بندی کننده ها، کاهش ابعاد، الگوریتم الگوریتم ماشین بردار پشتیبان ، الگوریتم پرسپترون چند لایه، الگوریتم نزدیکترین همسایه

مقدمه

در یادگیری ماشین و شناسایی الگو، طبقه بندی نوعی یادگیری با نظارت تلقی میشود که اشاره به نحوه حل مسئله تشخیص تعلق یک مشاهده جدید به یکی از کلاسها دارد. آموزش مدل با نظارت بر اساس مجموعه ای از داده هایی برچسب دار است که برچسب ها، تعلق مشاهدات به کلاس خاصی را تعیین میکند. بررسی و مقایسه ی عملکرد الگوریتم های مختلف طبقه بندی در مسائل کاربردی یادگیری ماشین و شناخت الگو، همواره یک چالش بوده است. همچنین امروزه با افزایش میزان داده های تولید شده، در به کار گیری الگوریتم های طبقه بندی کننده، با مشکل بالا بودن حجم محاسبات مواجه هستیم. به همین دلیل، کاربرد الگوریتم هایی که در پی کاهش ابعاد داده ها هستند، مورد توجه بسیاری قرار گرفته است. بطور خلاصه نتایج پژوهش های مهمی که در زمینه ارزیابی طبقه بندی کننده ها صورت پذیرفته اند، در زیر آورده شده اند.

لکون و همکارانش، پژوهشی انجام دادند که چند الگوریتم یادگیری(از جمله ماشین بردار پشتیبان) را روی مسائل تشخیص دستخط بر اساس دقت و هزینه محاسبه الگوریتم، مقایسه میکرد.[1]

کوپر و همکارانش نتایجی از یک پژوهش ارائه دادند که عملکرد ۱۲ الگوریتم یادگیری را روی داده های پزشکی، با استفاده از معیارهای دقت و ROC میسنجید.[2]

لیم و همکارانش یک مقایسه ی تجربی بین الگوریتم درخت تصمیم و سایر الگوریتم های طبقه بندی انجام دادند که آنها را بر اساس معیار دقت میسنجید.[3]

هدف از ارائه این مقاله، بررسی میزان پایداری و عملکرد چهار الگوریتم طبقه بندی کننده بیزین، نزدیکترین همسایه، ماشین بردار پشتیبان و پرسپترون چند لایه بر روی مجموعه داده IRIS در برابر کاهش ابعاد داده ها توسط الگوریتم PCA میباشد. در این مقاله جهت بررسی تاثیر کاهش بعد روی نتایج الگوریتم ها،داده ها را در دو مرحله وارد الگوریتم نموده و دقت آنها را ثبت نمودیم، در مرحله اول، مجموعه داده با تعداد ابعاد پیش فرض را وارد الگوریتم نمودیم، در مرحله ی بعد،ابعاد داده ها را کاهش دادیم، آنرا وارد الگوریتم نموده و نتایج را ثبت نمودیم. در نهایت مقایسه ای بین نتیجه عملکرد الگوریتم در برابر داده ها با ابعاد پیش فرض و ابعاد کاهش یافته انجام دادیم.

در ادامه این مقاله، در بخش دوم، به توضیح مدلها و روشهای به کار رفته خواهیم پرداخت. در بخش سوم به تشریح داده ها و آزمایشات انجام شده روی آنها و در بخش چهارم به جمع بندی و نتیجه گیری خواهیم پرداخت.

روش ها و مدل ها

از پرکاربردترین ابزارها جهت خودکار سازی فرایندها , و شناسایی الگوها میتوان به یادگیری ماشین اشاره کرد. الگوریتم هایی نظیر بیزین، نزدیکترین همسایه، ماشین بردار پشتیبان، پرسپترون چند لایه و ... که با دقت بسیاری الگوها را از داده ها و اطلاعات استخراج کرده و به سیستم میدهند، تا سیستم با چگونگی برخورد با الگوهای ورودی

مختلف در طول انجام فرایند آشنا شود، همگی به طور عمده به خودکار سازی فرآیند کمک میکنند.

در این بخش به معرفی این چهار الگوریتم میپردازیم:

الگوریتم ماشین بردار پشتیبان: الگوریتم بردار پشتیبان یا SVM، مجموعه ای از نقاط در فضای n بعدی داده ها هستند که مرز دسته ها را مشخص می کنند و دسته بندی داده ها براساس آنها انجام می شود و با جابجایی یکی از آنها، خروجی دسته بندی ممکن است تغییر کند. در فضای دوبعدی ،بردارهای پشتیبان، یک خط، در فضای سه بعدی یک صفحه و در فضای n بعدی یک ابر صفحه را شکل میدهد. در الگوریتم ماشین بردار پشتیبان فقط داده های قرار گرفته در بردارهای پشتیبان مبنای یادگیری ماشین و ساخت مدل قرار می گیرند و این الگوریتم به سایر نقاط داده حساس نیست و هدف آن هم یافتن بهترین مرز در بین داده هاست به گونه ای که بیشترین فاصله ممکن را از تمام دسته ها (بردارهای پشتیبان آنها) داشته باشد.

الگوریتم نزدیکترین همسایه: الگوریتم نزدیکترین همسایه در بسیاری از کاربرد های یادگیری ماشین ، همچون طبقه بندی، تشخیص ناهنجاری و ... مورد استفاده قرار میگیرد. الگوریتم K نزدیکترین همسایه یا KNN برای یک نمونه از داده ها مثل نقطه Q، از K همسایه آن رای گیری میکند و بر اساس بیشترین آرا داده شد به یک برچسب، داده مورد نظر را برچسب گذاری میکند. یک روش ساده برای پیدا کردن k نزدیکترین همسایه به نقطه Q نیاز به اسکن خطی تمام نقاط موجود در مجموعه M دارد. از آنجایی که این کار در عمل برای مجموعه داده های بزرگ و الگوریتم های تعیین فاصله مختلف، بسیار زمان بر و هزینه بر است، کاهش ابعاد داده یکی از راه حل های افزایش سرعت عملکرد این الگوریتم است که تاثیر آن روی نتیجه الگوریتم در این مقاله بررسی خواهد شد.

پرسپترون چند لایه: الگوریتم پرسپترون چند لایه(MLP)، شامل شبکه‌ای از پرسپترونها است، که می‌تواند رفتار پیچیده کلی تعیین شده‌ای از ارتباط بین عناصر پردازش و پارامترهای عنصر را نمایش دهد. این شبکه‌ها برای تخمین و تقریب، کارایی بسیار بالایی از خود نشان داده‌اند. کاربرد این مدل‌های ریاضی بر گرفته از عملکرد مغز انسان، بسیار وسیع می‌باشد که به عنوان چند نمونه کوچک می‌توان استفاده از این ابزار در پردازش سیگنال‌های الکترونیکی، مخابراتی و... را نام برد. ایده اصلی این روش، از سیستم مرکزی عصبی مغز انسان، نورونها، شاخه‌های متعدد سلولهای عصبی و محلهای تماس دو عصب نشأت گرفته‌است.

الگوریتم بیزین: این الگوریتم یک روش ساده برای طبقه بندی اطلاعات بر پایه ی احتمال وقوع یا عدم وقوع یک پدیده میباشد.

در داده ها با حجم و تعداد ویژگی های بالا، بار پردازش روی این داده ها نیز به مراتب زیاد میباشد. به جهت کمتر شدن بار محاسبات، از الگوریتم هایی نظیر PCA برای کاهش حجم داده استفاده میشود، این روش یک فضای چند بعدی را به فضایی با تعداد ابعاد کمتر نگاشت میکند. در واقع با ترکیب مقادیر ویژگیهای موجود، تعداد کمتری ویژگی بوجود می‌آورد بطوریکه این ویژگیها دارای تمام (یا بخش اعظمی از) اطلاعات موجود در ویژگیهای اولیه باشد. کاهش ابعاد داده با هدف افزایش سرعت الگوریتم، کاهش فضای ذخیره سازی، ترسیم و به دست آوردن درک کلی از داده ها، کاهش احتمال فرابرازش و در نتیجه آن افزایش تعمیم پذیری الگوریتم انجام میشود.

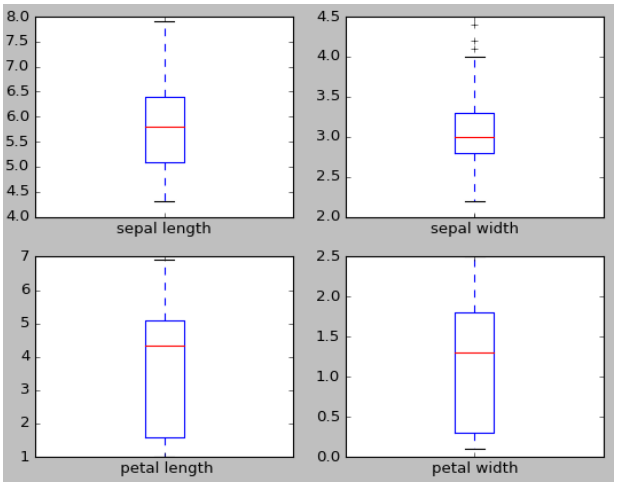
همانطور که در بخش قبل اشاره شد، در این مقاله قصد داریم تاثیر کاهش ابعد داده را روی نتایج چهار الگوریتم ذکر شده، بررسی کنیم. نتایج الگوریتم ها را با معیار Accuracy که نشان دهنده دقت الگوریتم میباشد، بیان میکنیم. این معیار توسط فرمول زیر محاسبه میشود:

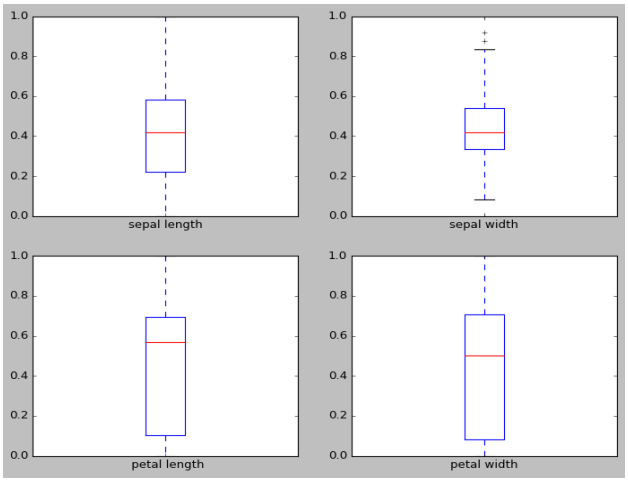
$$Accuracy = 1 - CER$$
$$CER = (Missclassification / AllData)$$

Missclassification : تعداد داده هایی که توسط الگوریتم، برچسب ناصحیح دریافت کرده اند و در طبقه ای دیگری طبقه بندی شده اند.

All data : تعداد تمام داده های مجموعه داده.

در طول انجام آزمایشات، از زبان برنامه نویسی پایتون و کتابخانه ی sklearn جهت مصور سازی، پیش پردازش و پیاده سازی الگوریتم ها استفاده نمودیم. پس از مصور سازی مجموعه داده، متوجه پراکندگی خصیصه های آن شدیم و به دلیل تاثیر منفی که پراکندگی داده ها بر روی الگوریتم های مورد استفاده بخصوص الگوریتم پرسپترون چند لایه دارد، جهت متراکم سازی داده ها اقدام نمودیم. در این گام پراکندگی ویژگی های مجموعه داده را به بازه ی ۰ تا ۱ نگاشت نمودیم. شکل ۱ (شکل بالا) نمایانگر بازه خصیصه های مجموعه داده بدون پیش پردازش و (شکل پایین) نمایانگر خصیصه ها پس از پردازش و نگاشت به بازه جدید ۰ تا ۱ میباشد.

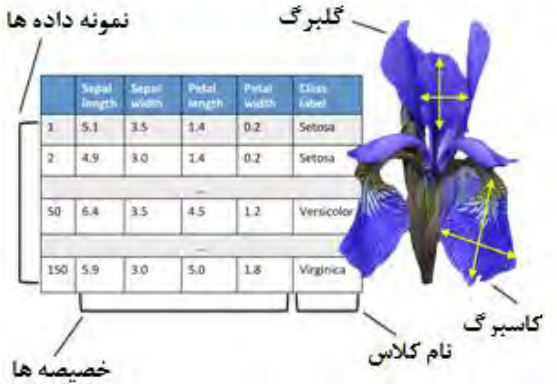




شکل۱- خصیصه های مجموعه داده قبل از پردازش(شکل بالا) و خصیصه های مجموعه داده پس از پردازش و تغییر بازه (شکل پایین)

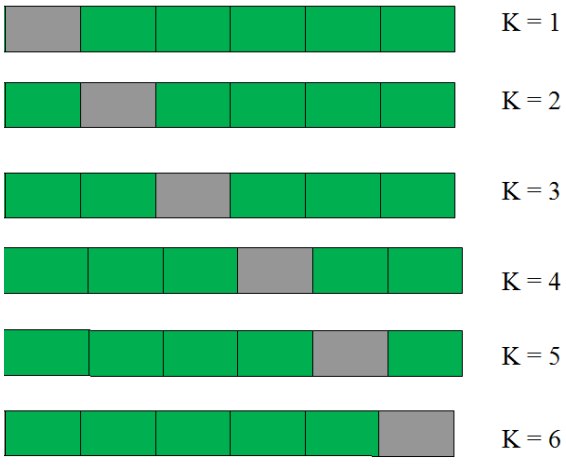
داده ها و آزمایش ها

در این آزمایش از مجموعه داده IRIS که یکی از مجموعه داده های موجود در سایت داتشگاه کالیفرنیا است ، استفاده نمودیم[1]. این نمونه داده متشکل از ۱۵۰ داده است که در سه کلاس مختلف تقسیم بندی شده اند و هر داده مربوط به یک گیاه است. هر نمونه از داده ها، متشکل از پنج مشخصه طول گلبرگ، پهنای گلبرگ، طول کاسبرگ، پهنای کاسبرگ و نام کلاس است. چهار ویژگی اول به وضوح بیانگر طول و پهنای برگ و کاسبرگ بر اساس سانتیمتر، و ویژگی پنجم نام کلاسی که نمونه به آن تعلق دارد را مشخص میکند. خصیصه پنجم سه کلاس با نامهای IRIS setosa، IRIS versicolours و IRIS viriginica را مشخص میکند. الگوریتم هایی که به کار میبریم با بهره بردن از چهار ویژگی اول به عنوان پارامترهای ورودی مدل سعی در تشخیص پارامتر پنجم (نوع کلاس) دارند. شکل ۲ نمایی مختصر از این مجموعه داده را نمایش میدهد.



شکل۲- نمای کلی مجموعه داده IRIS

به دلیل کم بودن تعداد داده ها، ما از اعتبار سنجی ضربدری (Cross-Validation) برای آموزش مدلها استفاده نمودیم. در این نوع اعتبارسنجی، داده ها به K زیرمجموعه افراز می شوند. از این K زیرمجموعه، هر بار یکی برای اعتبارسنجی و K-1 تای دیگر برای آموزش بکار میروند. این روال K بار تکرار می شود و همه داده ها دقیقا یکبار برای آموزش و یکبار برای اعتبارسنجی بکار می روند. در نهایت میانگین نتیجه این K بار اعتبارسنجی به عنوان یک تخمین نهایی برگزیده می شود. همانطوری که در شکل۳ ملاحظه میفرمایید، در هر ردیف، خانه های سبز رنگ مجموعه داده، داده های آموزش و خانه های خاکستری رنگ مجموعه داده به عنوان داده های تست به کار میروند. در نهایت الگوریتم به ازای هر ردیف یکبار اجرا میشود (مجموعا ۶ بار) و میانگین دقت ۶ بار اجرای الگوریتم به عنوان دقت الگوریتم در نظر گرفته میشود.



شکل ۳- مجموعه داده پس از اعمال اعتبارسنجی ضربدری

هر یک از الگوریتم های مورد آزمایش، دارای پارامترهای مختلفی میباشند که تغییر آنها در نتیجه ای که بدست می آورند، تاثیر گذار است. اینک توضیحی از پارمترهایی که الگوریتم ها در اختیار ما قرار میدهند، ارائه داده و سپس نتیجه اجرای الگوریتم با تغییر این پارامتر ها را بصورت جدولی نمایش خواهیم داد.

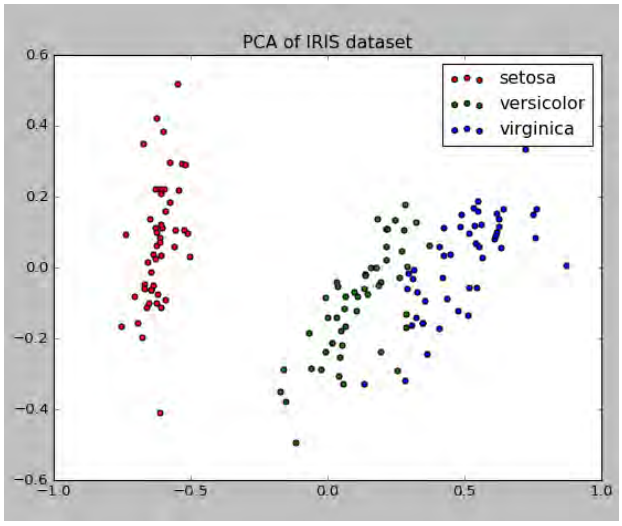
در الگوریتم پرسپترون چند لایه ، تنها یک لایه ی پنهان با تعداد نرون متغیر که آنرا در جدول با #hiddenLayer نمایش داده ایم را در نظر گرفتیم. همچنین مقدار نرخ یادگیری را برابر ۰،۳ و تعداد اپوکها را برابر ۲۰۰ در نظر گرفتیم.

الگوریتم k نزدیکترین همسایه تنها یک پارامتر را، که تعیین کننده تعداد همسایگی ها است، را در اختیار ما قرار میدهد؛ آنرا با K در جدول مشخص نمودیم.

الگوریتم ماشین بردار پشتیبان پارامترهای Kernel و C را در اختیار ما قرار میدهد که به ترتیب بیانگر هسته الگوریتم و میزان هموار بودن مرزهای جدا سازی نواحی کلاسها، میباشند.

الگوریتم بیزین تنها داده ها را به عنوان پارامتهای ورودی پذیرفته و هیچ پارامتر دیگری برای تغییر در اختیار ما قرار نمی دهد.

جهت بررسی تاثیر کاهش بعد توسط الگوریتم PCA روی نتیجه الگوریتم ها، هر داده را در دو مرحله وارد الگوریتم ها نموده و دقت آنها را ثبت نمودیم، در مرحله اول، داده های معمولی با تعداد ۴ بعد را وارد الگوریتم نمودیم، در مرحله ی بعد،ابعاد داده ها را توسط الگوریتم PCA از ۴ بعد به ۲ بعد کاهش دادیم و سپس آنرا وارد الگوریتم نمودیم. در شکل ۴، ساختار کلی مجموعه داده پس از کاهش بعد مشخص شده است.



شکل ۴- مجموعه داده مصور شده پس از اعمال PCA

در ادامه نتایج بدست آمده از اجرای الگوریتم ها با داده های معمولی و داده های کاهش بعد یافته را در جداول ۱ تا ۴ مشاهده می نمایید.

جدول ۱- نتایج اجرای الگوریتم پرسپترون چند لایه

#	hidden layer#	MLP Accuracy	MLP+PCA Accuracy
1	8	0.96	0.97
3	20	0.96	0.95
3	50	0.95	0.96
4	111	0.96	0.96

جدول ۲- نتایج اجرای الگوریتم نزدیکترین همسایه

K	KNN Accuracy	KNN+PCA Accuracy
1	0.96	0.96
3	0.97	0.97
5	0.97	0.96
7	0.96	0.95
11	0.98	0.95

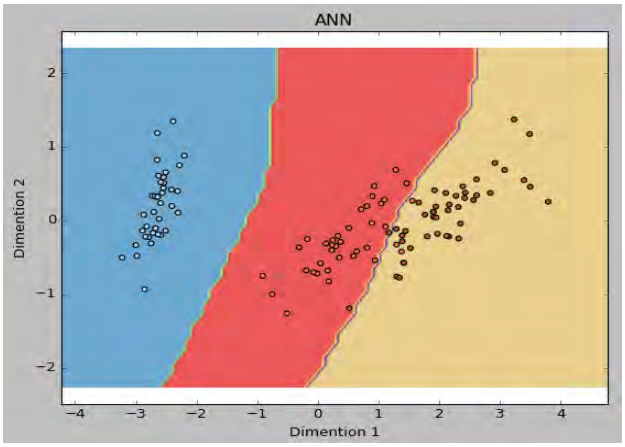
جدول ۳- نتایج اجرای الگوریتم ماشین بردار پشتیبان

Kernel	Parameter C	SVM Accuracy	SVM+PCA Accuracy
Linear	0.01	0.93	0.91
Linear	0.1	0.97	0.95
Linear	1	0.98	0.95
Poly	0.01	0.95	0.64
Poly	0.1	0.98	0.73
Poly	1	0.97	0.87

جدول۴- نتایج اجرای الگوریتم بیزین

#	NB Accuracy	NB + PCA Accuracy
1	0.95	0.89

بعنوان نمونه ای از عملکرد الگوریتم ها به صورت مصور شده، در شکل ۵، نحوه ی ناحیه بندی الگوریتم پرسپترون چند لایه روی داده های کاهش بعد یافته را مشاهده مینمایید.



شکل۵- ناحیه های تفکیک شده توسط الگوریتم پرسپترون چندلایه

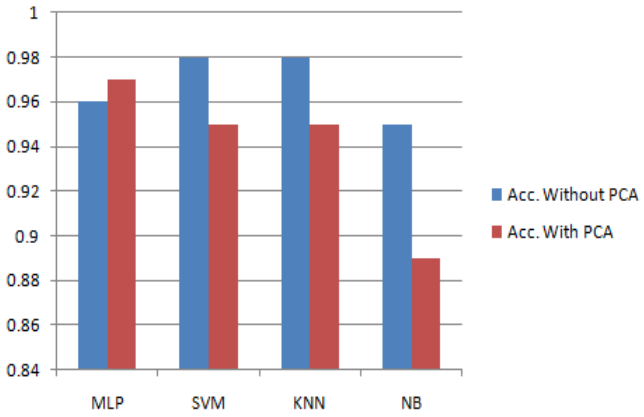
جمع بندی و نتیجه گیری

همانطور که در بخش قبل مشاهده نمودید، الگوریتم ها علاوه بر داده ها، پارامترهای دیگری را بعنوان ورودی دریافت میکنند تا انعطافی نسبی در برابر داده های ورودی از خود نشان دهند. با تغییر این پارامترها توانستیم چندین جواب از یک الگوریتم بدست آوریم، اما جوابی که بیشترین دقت را داشت، به عنوان خروجی الگوریتم برگزیدیم. نتایج الگوریتم ها در جدول ۵ آمده است.

جدول ۵- نتیجه ی اجرای الگوریتم ها

	Accuracy without PCA	Accuracy with PCA
MLP	0.96	0.97
SVM	0.98	0.95
KNN	0.98	0.95
NB	0.95	0.89

همانگونه که در شکل ۶ مشخص است، کاهش ابعاد داده بر روی الگوریتم پرسپترون چند لایه، تاثیر مثبت داشته است و باعث بهبود عملکرد این الگوریتم شده است. اما نتایج بدست آمده از سایر الگوریتم ها، نشان میدهد این الگوریتم ها در برابر کاهش ابعاد داده، عملکرد ضعیف تری دارند.



شکل6- تاثیر کاهش داده بر روی الگوریتم ها

منابع

[1] King RD, Feng C, Sutherland A, 1995. “comparison of classification algorithms on large real-world problems.” Applied Artificial Intelligence an International Journal.,9(3),May , pp.289-333.

[2] Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, Fine MJ, Glymour C, Gordon G, Hanusa BH, Janosky JE.1997. An evaluation of machine-learning methods for predicting pneumonia mortality. Artificial intelligence in medicine, , 9(2),Feb, pp.107-138.

[3] Lim TS, Loh WY, Shih YS, 2000” A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms.” Machine learning, *40*(3), Sep, pp.203-228.

[4]