دانشگاه تهران

دانشکده فنی-مهندسی کامپیوتر

دپارتمان الگوریتم ها و محاسبات

گزارش تمرین شماره ی دو

طراحی طبقه بند

نیلوفر آقایی ابیانه

۸۱۰۸۹۰۰۰۱

چکیده

در این پروژه ، الگوریتم های مختلف طبقه بند[۱] روی مجموعه های داده ای متعدد اجرا می شود و نتایج الگوریتم های تحلیل می شود. برای این کار از الگوریتم ***های طبقه بند اولین نزدیک ترین همسایه***[۲]، ***طبقه بند بیز***[۳]، ***k*مین نزدیک ترین همسایه**[۴] **و پنجره ی پارزن** [۵]استفاده می شود.

۱. مقدمه

در علم کامپیوتر، الگوریتم های مختلفی برای طبقه بندی وجود دارد، که بر اساس انجام کار به سه دسته ی اصلی تقسیم می شوند:

i. ***طبقه بندها بر اساس مفهوم شباهت***[۶]
ii. ***طبقه بندها بر اساس روش های احتمالی***[۷]
iii ***طبقه بندها بر اساس ساخت مرز تصمیم گیری توسط بهینه سازی معیار های خطا***

در این گزارش از پنج مجموعه های داده ای[۸] تحت نام های: *مجموعه ای داده ای_۱ ، مجموعه ای داده ای_۲، مجموعه ای داده ای_phoneme، مجموعه ای داده ای_iris* و *مجموعه ای داده ای_satimage* استفاده شده و روی الگوریتم های طبقه بند *اولین نزدیک ترین همسایه، طبقه بند بیز، k*مین نزدیک ترین *همسایه و پنجره ی پارزن* اجرا می شوند.

اندازه ی مجموعه ای داده ای_۱ ، مجموعه های داده ای_۲ در هر بار اجرای الگوریتم ها متغییر است در حالیکه اندازه ی مجموعه های داده ای_phoneme، مجموعه های داده ای_iris و مجموعه های داده ای_satimage ثابت می باشند. در واقع از مجموعه ای داده ای_۱ ، مجموعه های داده ای_۲، تعدادی نقطه، به صورت تصادفی برای هر الگوریتم تولید شده و در آخر نتایج به دست آمده با هم مقایسه می شوند.

---

۱ Classifier
۲ One-nearest Neighbor
۳ Bayes Classifier
۴ K-nearest Neighbor
۵ Parzen Window
۶ Concept of Similarity
۷ Probabilistic Approach
۸ Dataset

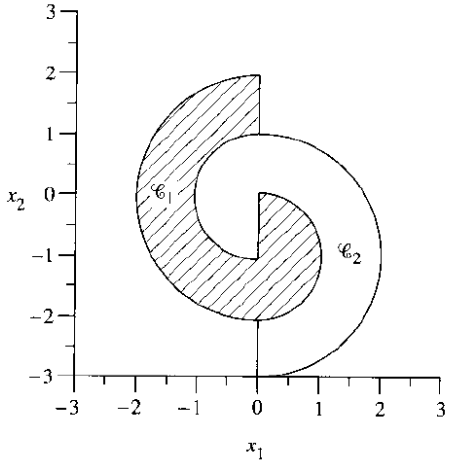
برای پیاده سازی این الگوریتم ها و همچنین انتخاب نقاط از مجموعه ای داده ای_۱ ، مجموعه های داده ای_۲ از محیط MATLAB استفاده می شود.

۲.مجموعه های داده ای

همانطور که گفته شد در این پروژه از پنج مجموعه ی داده ای استفاده می شود.

i. مجموعه ای داده ای_۱

نقاط این مجموعه های داده ای در هر بار آزمایش از بازه ای که در شکل ۱ نشان داده شده است، انتخاب می شوند.

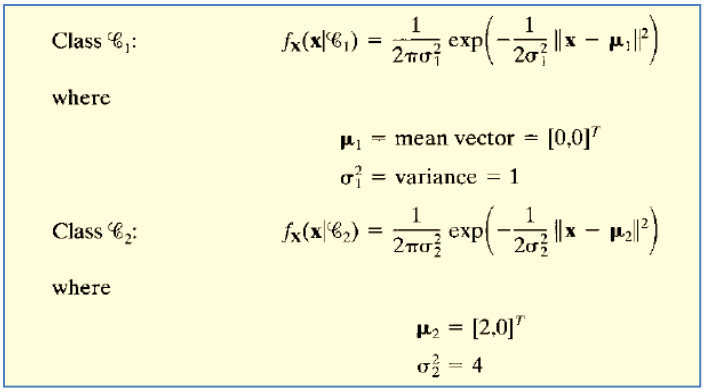


شکل۱-بازه ی انتخاب نقاط

برای این کار در محیط MATLAB تابع dataset_۱ تعریف شده است؛ این تابع در هر بار اجرا، جداگانه برای هر کلاس با استفاده از توزیع یکنواخت ، به اندازه ی pointnumber نقطه انتخاب می کند و در ماتریس های class_۱و class_۲ قرار می دهد. به ماتریس class_۱ برچسب A و به ماتریس class_۲ بر چسب B می زند . در نتیجه، یک مجموعه های داده ای با دو طبقه بدست می آید.

ii. مجموعه های داده ای_۲

نقاط این مجموعه های داده ای در هر بار آزمایش از بازه ای که درفرمول شکل ۲ نشان داده شده است، انتخاب می شوند.



شکل۲-فرمول بدست آوردن نقاط با استفاده از تابع گاوس

برای این کار در محیط MATLAB تابع dataset_۲ تعریف شده است؛ این تابع در هر بار اجرا، جداگانه برای هر کلاس با استفاده از توزیع نرمال ، به اندازه ی numberofpoint نقطه انتخاب می کند و در ماتریس های class_۱و class_۲ قرار می دهد. به ماتریس class_۱ برچسب A و به ماتریس class_۲ بر چسب B می زند . در نتیجه، یک مجموعه های داده ای با دو طبقه بدست می آید.

iii. مجموعه های داده ای_phoneme

نقاط این مجموعه ای داده ای از پایگاه داده ای ELENA تحت عنوان phoneme بدست آمده است[۱] . این مجموعه ای داده ای متشکل از سه ماتریس به اندازه ی ۵۴۰x۶ است.

iv. مجموعه های داده ای_iris

نقاط این مجموعه ای داده ای از پایگاه داده ای ELENA تحت عنوان Iris بدست آمده است[۲] . این مجموعه ای داده ای، متشکل از یک ماتریس به اندازه ی ۱۵۰x۵ است.

v. مجموعه ای داده ای_satimage

نقاط این مجموعه ای داده ای از پایگاه داده ای ELENA تحت عنوان satimage بدست آمده است[۳] . این مجموعه ای داده ای، متشکل از سه ماتریس به اندازه ی ۶۴۳۵x۳۷ است.

۳.الگوریتم ها

همانطور که گفته شد در این پروژه از سه مجموعه ی داده ای استفاده می شود؛ که در زیر آمده است.

I. طبقه بند اولین نزدیک ترین همسایه (OneNN)

در این روش فاصله ی هر نمومه تست، با همه ی نمونه های train مقایسه می شود و در آخر بر چسب نزدیکترین نقطه را به خود می گیرد.

برای این الگوریتم، در MATLAB تا بع OneNN تعریف شده است. این تابع دو ماتریس train و test و شماره ی مجموعه ای داده ای را به عنوان ورودی می گیرد و حاصل را در خروجی را در ماتریس matrix قرار می دهد.

II. طبقه بند بیز(Bayes)

در این روش تعلق یک نمونه تست، بر اساس تابع احتمال بیز (فرمول زیر) مشخص می شود.

$$p(wi|x) = \frac{p(x|wi)p(wi)}{p(x)}$$

که پس از محاسبات تعلق یک نمونه تست با استفاده از فرمول شکل ۳ بدست می آید.

$$g_i(x) = P(\omega_i \mid x)$$
$$= \mathrm{P(x} \mid \omega_i)\,\mathrm{P}(\omega_i)$$
... or equivalent ly
$$= \log P(x \mid \omega_i) + \log P(\omega_i)$$
if we can assume that $P(x \mid \omega_i)$ *are Gaussian*
$$P(x \mid \omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$
$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(\omega_i)$$
for multivariate:
$$P(x \mid \omega_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right]$$ d: input dimension
$$g_i(x) = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \log P(\omega_i)$$

شکل۳-فرمول بیز برای تعلق یک عنصر به یک کلاس

برای این الگوریتم، در MATLAB تا بع Bayes تعریف شده است. این تابع دو ماتریس train ، test و شماره ی مجموعه ای داده ای را به عنوان ورودی می گیرد و حاصل را در خروجی را در ماتریس matrix قرار می دهد.

III. طبقه بند kمین نزدیک ترین همسایه (KNN)

این روش مشابه روش اولین نزدیک ترین همسایه می باشد. در این روش فاصله ی هر نمومه تست، با همه ی نمونه های train مقایسه می شود و در آخرkتا نزدیکترین عنصر به نمونه تست در نظر گرفته می شوند و بر چسب نمونه تست برچسب بیشترین را می گیرد اما در شرایطی که برابر باشد، برای بر چسب نمومه تست، روش های متعددی وجود دارد.

در این پروژه، برای بر چسب نمومه تست، میانگین فاصله kتا عنصر محاصبه می شود؛ سپس بر چسب عنصری که به میانگین نزدیکتر است به عنوان بر چسب نمومه تست در نظر گرفته می شود.

برای این الگوریتم، در MATLAB تا بع KNN تعریف شده است. این تابع دو ماتریس train ، test ، شماره ی مجموعه ای داده ای و K را به عنوان ورودی می گیرد و حاصل را در خروجی را در ماتریس matrix قرار می دهد.

در اینجا هر پنج مجموعه ای داده ای با دو مقدار K=۵ و K=۱۰ بررسی می شوند.

IV. طبقه بند پنجره ی پارزن(ParzonWindows)

در این روش تعلق یک نمونه تست، بر اساس تابع احتمال زیر مشخص می شود.

$$p(x) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Ni}{h}\varphi\left(\frac{ci-x}{h}\right)\right)$$

در واقع یک شعاع ثابت،H، در نظر گرفته می شود؛ که بر اساس آن برای نمونه تست آمده شده بر چسب نمونه تست انتخاب می شود.

برای این الگوریتم، در MATLAB تا بع ParzonWindows تعریف شده است. این تابع دو ماتریس train ، test ، شماره ی مجموعه ای داده ای ، H را به عنوان ورودی می گیرد و حاصل را در خروجی را در ماتریس matrix قرار می دهد.

H متغیری است که به صورت تجربی بدست می آید. بنا به ویژگی مجموعه های داده ای مقدار H برای هر کدام متفاوت است.

در اینجا برای هر مجموعه های داده ای،هر الگوریتم برای دو مقدار متعدد H اجرا می شود. برای مجموعه های داده ای ۱ و۲ مقدار H=۰.۵ و H=۱ در نظر گرفته می شود در

حالیکه برای مجموعه های داده ای phoneme ، iris و satimage به ترتیب H=۱و H=۲ ، H=۰.۸ و H=۱.۷ و H=۷۰ و H=۱۸۰ انتخاب می شوند.
همانطور که مشاهده شد ورودی همه ی این توابع دو ماتریس train و test است. برای ایجاد این دو ماتریس در محیط MATLAB تابع selectingtestandtrain تعریف شده است؛ که این تابع ماتریس مجموعه ای داده ای را بر اساس درصدی (percentage) که از کاربر می گیرد به دو ماتریس trainو test تقسیم می کند.

برای اجرای این توابع با مجموعه های داده ای در MATLAB از تابع MAIN استفاده می شود؛ که در آن به ازای هر الگوریتم دستور switch-case به کار رفته است؛ به طوری که case۱ ، case۲ ، case۳ و case۴ به ترتیب به الگوریتم های OneNN ، Bayes ، KNN و ParzenWindows اشاره می کنند. از طرفی برای هر یک از این case ها به ازای هر مجموعه ای داده ای از دستور switch-case وجود دارد، به روشی که case۱ ، case۲ ، case۳ ، case۴ و case۵ به ترتیب به مجموعه های داده ای مجموعه ای داده ای_۱ ، مجموعه ای داده ای_۲، مجموعه ای داده ای_phoneme، مجموعه ای داده ای_iris و مجموعه ای داده ای_satimage دلالت می کنند.

این برنامه یکبار به ازای یک الگوریتم و مجموعه ای داده ای خاص که توسط کاربر مشخص می شود ، اجرا می گردد و نتیجه را به کاربر اعلام می کند. برای اجرای بعدی تابع MAIN مجددا باید فراخوانی شود.

در این برنامه از توابع زیر استفاده شده است:

- MAIN : تابع اصلی
- dataset_۱ : تولید اعداد مجموعه ای داده ای_۱ به تعداد pointnumber
- dataset_۲: تولید اعداد مجموعه ای داده ای_۲ به تعداد numberofpoint
- OneNN : تابع اجرایی الگوریتم طبقه بند اولین نزدیک ترین همسایه
- Bayes : تابع اجرایی الگوریتم طبقه بند بیز
- KNN : تابع اجرایی الگوریتم طبقه بند kامین نزدیک ترین همسایه
- ParzenWindows: تابع اجرایی الگوریتم طبقه بند پنجره ی پارزن

- Repmatman : گسترش بردار ورودی به ابعاد خواسته شده
- removeonedimension : حذف بعد اول ماتریس ورودی
- insertionsort : مرتب سازی ماتریس ورودی بر اساس یک ستون خاص
- findclasses : بر گرداندن لیست همه ی کلاس ها ی trainingset
- selectingtestandtrain : انتخاب ماتریس های train و test با درصد مورد نظر کاربر
- insertionsoret_onlabel : مرتب سازی ماتریس ورودی بر اساس کلاس ها
- efficiency : محاصبه ی درستی طبقه بندها (عددی بین ۰ -۱)

۴.آزمایش ها و نتایج

در این قسمت نتایج حاصل از اجرای الگوریتم های مختلف روی مجموعه های داده ای متفاوت با اندازه های متعدد در جداول ۱-۲۰ آمده است؛ که در هر جدول هر سطر بیانگر اندازه ی مجموعه های داده ای و هر ستون بیانگر درصد انتخاب عناصر تست و هر داریه جدول بیانگر کارایی آن الگوریتم روی اندازه ی مجموعه های داده ای خاص با درصد مشخص است.

أ. با انجام آزمایش ها روی الگوریتم OneNN و مجموعه های داده ای: مجموعه ای داده ای_۱ ، مجموعه ای داده ای_۲، مجموعه ای داده ای_phoneme، مجموعه ای داده ای_iris و مجموعه ای داده ای_satimage نتایج حاصل در جدول های ۱-۵ آمده است.

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۱۰۰x۳	۰.۹۰۰۰	۰.۹۰۰۰	۰.۷۶۶۷	۰.۸۲۵۰	۰.۸۱۶۷
۲۰۰x۳	۰.۶۵۰۰	۰.۷۵۰۰	۰.۸۰۰۰	۰.۸۰۰۰	۰.۸۴۱۷
۵۰۰x۳	۰.۸۴۰۰	۰.۸۰۰۰	۰.۸۵۳۳	۰.۸۵۵۰	۰.۸۷۶۷
۱۰۰۰x۳	۰.۸۷۰۰	۰.۸۴۵۰	۰.۷۹۶۷	۰.۸۲۷۵	۰.۸۱۳۳
۲۰۰۰x۳	۰.۸۳۰۰	۰.۷۹۷۵	۰.۸۴۳۳	۰.۸۴۵۰	۰.۸۱۰۸

جدول ۱-نتایج حاصل از اجرای الگوریتم OneNN روی مجموعه ای داده ای_۱ با درصد تست های مختلف

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۱۰۰x۳	۰.۹۰۰۰	۰.۹۵۰۰	۰.۹۰۰۰	۰.۸۲۵۰	۰.۸۱۶۷
۲۰۰۰x۳	۰.۸۵۰۰	۰.۸۰۰۰	۰.۸۰۰۰	۰.۹۰۰۰	۰.۹۰۰۰
۵۰۰x۳	۰.۸۶۰۰	۰.۸۲۰۰	۰.۸۴۶۷	۰.۸۳۰۰	۰.۸۶۰۰
۱۰۰۰x۳	۰.۷۹۰۰	۰.۸۶۰۰	۰.۸۱۰۰	۰.۸۳۷۵	۰.۸۳۰۰
۲۰۰۰x۳	۰.۸۵۵۰	۰.۸۶۷۵	۰.۸۶۵۰	۰.۸۶۱۷	۰.۸۵۸۳

جدول ۲-نتایج حاصل از اجرای الگوریتم OneNN روی مجموعه ای داده ای_۲ با درصد تست های مختلف

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۵۴۰۴x۶	۰.۹۰۰۴	۰.۹۰۷۶	۰.۸۹۶۴	۰.۸۹۹۲	۰.۹۱۱۲

جدول ۳-نتایج حاصل از اجرای الگوریتم OneNN روی مجموعه ای داده ای_phoneme با درصد تست های مختلف

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۱۵۰x۵	۱	۱	۰.۹۵۶۵	۰.۹۳۳۳	۱

جدول ۴-نتایج حاصل از اجرای الگوریتم OneNN روی مجموعه ای داده ای_iris با درصد تست های مختلف

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۶۴۳۵x۳۷	۰.۹۰۳۷	۰.۸۸۹۸	۰.۸۸۵۱	۰.۹۰۰۵	۰.۹۰۰۶

جدول ۵-نتایج حاصل از اجرای الگوریتم OneNN روی مجموعه ای داده ای_satimage با درصد تست های مختلف

ب. با انجام آزمایش ها روی الگوریتم Bayes و مجموعه های داده ای: مجموعه ای داده ای_۱ ، مجموعه ای داده ای_۲، مجموعه ای داده ای_phoneme، مجموعه ای داده ای_iris و مجموعه ای داده ای_satimage نتایج حاصل در جدول های ۶-۱۰ آمده است.

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۱۰۰x۳	۰.۹۰۰۰	۰.۸۰۰۰	۰.۶۷۶۷	۰.۶۳۵۰	۰.۸۰۰۰
۲۰۰x۳	۰.۸۰۰۰	۰.۸۷۵۰	۰.۷۸۳۳	۰.۸۱۲۵	۰.۷۹۱۷
۵۰۰x۳	۰.۶۴۰۰	۰.۷۵۰۰	۰.۷۲۰۰	۰.۸۰۰۰	۰.۷۳۶۷
۱۰۰۰x۳	۰.۷۳۰۰	۰.۷۰۰۰	۰.۷۳۶۷	۰.۷۲۷۵	۰.۶۹۶۷
۲۰۰۰x۳	۰.۷۲۵۰	۰.۷۳۲۵	۰.۷۳۶۷	۰.۷۳۲۵	۰.۷۱۴۲

جدول ۶-نتایج حاصل از اجرای الگوریتم Bayes روی مجموعه ای داده ای_۱ با درصد تست های مختلف

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۱۰۰x۳	۰.۹۰۰۰	۰.۹۰۰۰	۰.۸۶۶۷	۰.۹۰۰۰	۰.۹۰۰۰
۲۰۰x۳	۰.۸۵۰۰	۰.۹۰۰۰	۰.۹۳۳۳	۰.۹۰۰۰	۰.۹۳۳۳
۵۰۰x۳	۰.۹۲۰۰	۰.۹۱۰۰	۰.۹۲۰۰	۰.۹۴۰۰	۰.۹۱۶۷
۱۰۰۰x۳	۰.۹۶۰۰	۰.۹۳۵۰	۰.۹۴۳۳	۰.۹۱۷۵	۰.۹۱۳۳
۲۰۰۰x۳	۰.۹۳۰۰	۰.۹۱۷۵	۰.۹۲۰۰	۰.۹۲۲۵	۰.۹۱۰۰

جدول ۷-نتایج حاصل از اجرای الگوریتم Bayes روی مجموعه ای داده ای_۲ با درصد تست های مختلف

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۵۴۰۴x۶	۰.۷۸۹۷	۰.۷۹۱۱	۰.۸۱۰۱	۰.۷۹۸۳	۰.۷۸۴۸

جدول ۸-نتایج حاصل از اجرای الگوریتم Bayes روی مجموعه ای داده ای_phoneme با درصد تست های مختلف

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۱۵۰x۵	۰.۸۷۵۰	۰.۹۳۳۳	۰.۹۵۶۵	۰.۹۶۶۷	۰.۹۵۵۶

جدول ۹-نتایج حاصل از اجرای الگوریتم Bayes روی مجموعه ای داده ای_iris با درصد تست های مختلف

	۰.۰۵	۰.۱	۰.۱۵	۰.۲	۰.۳
۶۵۴۳x۳۷	۰.۸۴۷۸	۰.۸۴۷۸	۰.۸۵۹۲	۰.۸۵۷۰	۰.۸۴۹۸

جدول ۱۰-نتایج حاصل از اجرای الگوریتم Bayes روی مجموعه ای داده ای_satimage با درصد تست های مختلف

ت. با انجام آزمایش ها روی الگوریتم KNN و انتخاب K=۵ و K=۱۰ روی مجموعه های داده ای: مجموعه ای داده ای_۱ ، مجموعه ای داده ای_۲، مجموعه ای داده ای_phoneme، مجموعه ای داده ای_iris و مجموعه ای داده ای_satimage نتایج حاصل در جدول های ۱۰-۱۵ آمده است.

	۰.۰۵ K=۵	۰.۰۵ K=۱۰	۰.۱ K=۵	۰.۱ K=۱۰	۰.۱۵ K=۵	۰.۱۵ K=۱۰	۰.۲ K=۵	۰.۲ K=۱۰	۰.۳ K=۵	۰.۳ K=۱۰
۱۰۰x۳	۱	۱	۰.۹۰۰۰	۰.۹۵۰۰	۰.۷۶۶۷	۰.۸۳۳۳	۰.۷۵۰۰	۰.۷۲۵۰	۰.۸۱۶۷	۰.۷۵۰۰
۲۰۰x۳	۰.۸۰۰۰	۰.۷۵۰۰	۰.۸۰۰۰	۰.۸۲۵۰	۰.۸۶۶۷	۰.۹۰۰۰	۰.۸۸۷۵	۰.۸۶۲۵	۰.۸۵۸۳	۰.۸۵۰۰
۵۰۰x۳	۰.۸۶۰۰	۰.۸۶۰۰	۰.۸۱۰۰	۰.۸۶۰۰	۰.۸۸۰۰	۰.۸۴۰۰	۰.۸۰۵۰	۰.۸۱۰۰	۰.۷۶۰۰	۰.۷۹۰۰
۱۰۰۰x۳	۰.۸۸۰۰	۰.۸۶۰۰	۰.۸۸۰۰	۰.۸۵۵۰	۰.۸۳۶۷	۰.۸۶۶۷	۰.۸۲۷۵	۰.۸۳۰۰	۰.۸۳۱۷	۰.۸۴۳۳
۲۰۰۰x۳	۰.۸۳۵۰	۰.۸۵۰۰	۰.۸۰۰۰	۰.۸۰۰۰	۰.۸۲۳۳	۰.۸۱۸۳	۰.۸۱۷۵	۰.۸۰۷۵	۰.۸۲۵۰	۰.۸۲۸۳

جدول ۱۱-نتایج حاصل از اجرای الگوریتم KNN روی مجموعه ای داده ای_۱ با درصد تست های مختلف به ازای K = ۵ و K = ۱۰ روی مجموعه های تست و train یکسان

	۰.۰۵ K=۵	۰.۰۵ K=۱۰	۰.۱ K=۵	۰.۱ K=۱۰	۰.۱۵ K=۵	۰.۱۵ K=۱۰	۰.۲ K=۵	۰.۲ K=۱۰	۰.۳ K=۵	۰.۳ K=۱۰
۱۰۰x۳	۰.۸۰۰۰	۰.۸۰۰۰	۰.۹۵۰۰	۰.۹۰۰۰	۰.۷۰۰۰	۰.۶۶۶۷	۰.۸۲۵۰	۰.۷۷۵۰	۰.۸۵۰۰	۰.۷۸۳۳
۲۰۰x۳	۰.۹۰۰۰	۰.۹۰۰۰	۰.۸۷۵۰	۰.۹۰۰۰	۰.۸۶۶۷	۰.۸۵۰۰	۰.۸۱۲۵	۰.۸۱۲۵	۰.۸۷۵۰	۰.۸۹۱۷
۵۰۰x۳	۰.۸۴۰۰	۰.۸۶۰۰	۰.۹۱۰۰	۰.۹۱۰۰	۰.۸۸۰۰	۰.۸۷۳۳	۰.۸۶۰۰	۰.۸۷۵۰	۰.۸۶۳۳	۰.۸۹۰۰
۱۰۰۰x۳	۰.۸۸۰۰	۰.۹۰۰۰	۰.۹۰۰۰	۰.۹۱۵۰	۰.۹۰۰۰	۰.۹۰۶۷	۰.۸۹۰۰	۰.۹۰۷۵	۰.۹۱۶۷	۰.۹۲۰۰
۲۰۰۰x۳	۰.۸۸۰۰	۰.۸۹۰۰	۰.۸۸۰۰	۰.۸۹۰۰	۰.۹۱۵۰	۰.۹۱۶۷	۰.۸۹۳۸	۰.۸۹۵۰	۰.۸۹۶۷	۰.۹۰۱۷

جدول ۱۲-نتایج حاصل از اجرای الگوریتم KNN روی مجموعه ای داده ای_۲ با درصد تست های مختلف به ازای K = ۵ و K = ۱۰ روی مجموعه های تست و train یکسان

	۰.۰۵ k=۵	۰.۰۵ k=۱۰	۰.۱ k=۵	۰.۱ k=۱۰	۰.۱۵ k=۵	۰.۱۵ k=۱۰	۰.۲ k=۵	۰.۲ k=۱۰	۰.۳ k=۵	۰.۳ k=۱۰
۵۴۰۴x۶	۰.۹۰۴۱	۰.۸۷۸۲	۰.۸۷۰۶	۰.۸۶۳۲	۰.۸۸۶۶	۰.۸۵۷۰	۰.۸۹۲۷	۰.۸۷۳۳	۰.۸۷۵۵	۰.۸۵۴۵

جدول ۱۳-نتایج حاصل از اجرای الگوریتم KNN روی مجموعه ای داده ای_phoneme با درصد تست های مختلف به ازای K =۵ و K =۱۰ روی مجموعه های تست و train یکسان

	۰.۰۵ k=۵	۰.۰۵ k=۱۰	۰.۱ k=۵	۰.۱ k=۱۰	۰.۱۵ k=۵	۰.۱۵ k=۱۰	۰.۲ k=۵	۰.۲ k=۱۰	۰.۳ k=۵	۰.۳ k=۱۰
۱۵۰x۵	۱	۱	۰.۹۳۳۳	۰.۸۶۶۷	۱	۰.۹۵۶۵	۰.۹۶۶۷	۱	۰.۹۷۷۸	۰.۹۷۷۸

جدول ۱۴-نتایج حاصل از اجرای الگوریتم KNN روی مجموعه ای داده ای_iris با درصد تست های مختلف به ازای K =۵ و K =۱۰ روی مجموعه های تست و train یکسان

	۰.۰۵ k=۵	۰.۰۵ k=۱۰	۰.۱ k=۵	۰.۱ k=۱۰	۰.۱۵ k=۵	۰.۱۵ k=۱۰	۰.۲ k=۵	۰.۲ k=۱۰	۰.۳ k=۵	۰.۳ k=۱۰
۶۴۳۵x۳۷	۰.۸۹۱۳	۰.۸۷۸۸	۰.۹۰۹۸	۰.۸۸۹۷	۰.۸۹۶۴	۰.۸۹۶۴	۰.۸۹۵۸	۰.۸۸۴۲	۰.۹۱۵۰	۰.۸۹۹۰

جدول ۱۵-نتایج حاصل از اجرای الگوریتم KNN روی مجموعه ای داده ای_satimage با درصد تست های مختلف به ازای K =۵ و K =۱۰ روی مجموعه های تست و train یکسان

ث. با انجام آزمایش ها روی الگوریتم ParzenWindows و مجموعه های داده ای: مجموعه ای داده ای_۱ ، مجموعه ای داده ای_۲، مجموعه ای داده ای_phoneme، مجموعه ای داده ای_iris و مجموعه ای داده ای_satimage نتایج حاصل در جدول های ۱۶-۲۰ آمده است.

	۰.۰۵ H=۰.۵	۰.۰۵ H=۱	۰.۱ H=۰.۵	۰.۱ H=۱	۰.۱۵ H=۰.۵	۰.۱۵ H=۱	۰.۲ H=۰.۵	۰.۲ H=۱	۰.۳ H=۰.۵	۰.۳ H=۱
۱۰۰x۳	۱	۰.۹۰۰۰	۰.۷۵۰۰	۰.۷۵۰۰	۰.۸۶۶۷	۰.۸۳۳۳	۰.۸۲۵۰	۰.۸۰۰۰	۰.۷۶۶۷	۰.۸۳۳۳
۲۰۰x۳	۰.۸۵۰۰	۰.۸۰۰۰	۰.۷۲۵۰	۰.۶۷۵۰	۰.۸۳۳۳	۰.۷۶۶۷	۰.۸۷۵۰	۰.۸۵۰۰	۰.۸۳۳۳	۰.۸۵۸۳
۵۰۰x۳	۰.۷۸۰۰	۰.۷۶۰۰	۰.۸۷۰۰	۰.۸۰۰۰	۰.۸۶۰۰	۰.۸۰۶۷	۰.۸۱۰۰	۰.۸۰۰۰	۰.۷۹۰۰	۰.۷۳۰۰
۱۰۰۰x۳	۰.۸۱۰۰	۰.۷۵۰۰	۰.۸۱۵۰	۰.۷۵۰۰	۰.۸۴۰۰	۰.۸۲۳۳	۰.۸۴۲۵	۰.۷۸۰۰	۰.۸۳۱۷	۰.۸۰۱۷
۲۰۰۰x۳	۰.۸۵۵۰	۰.۸۰۰۰	۰.۸۰۷۵	۰.۷۸۲۵	۰.۸۶۱۷	۰.۸۲۰۰	۰.۸۲۳۷	۰.۷۸۰۰	۰.۸۲۳۳	۰.۷۸۳۳

جدول ۱۶-نتایج حاصل از اجرای الگوریتم ParzenWindows روی مجموعه ای داده ای_۱ با درصد تست های مختلف و H=۰.۵ و H=۱ روی مجموعه های تست و train یکسان

	۰.۰۵ H=۰.۵	۰.۰۵ H=۱	۰.۱ H=۰.۵	۰.۱ H=۱	۰.۱۵ H=۰.۵	۰.۱۵ H=۱	۰.۲ H=۰.۵	۰.۲ H=۱	۰.۳ H=۰.۵	۰.۳ H=۱
۱۰۰x۳	۰.۵۰۰۰	۰.۶۰۰۰	۰.۵۰۰۰	۰.۸۰۰۰	۰.۴۰۰۰	۰.۶۰۰۰	۰.۴۰۰۰	۰.۶۰۰۰	۰.۵۳۳۳	۰.۷۳۳۳
۲۰۰x۳	۰.۷۵۰۰	۰.۸۵۰۰	۰.۶۰۰۰	۰.۸۰۰۰	۰.۷۱۶۷	۰.۸۰۰۰	۰.۶۵۰۰	۰.۷۷۵۰	۰.۶۰۰۰	۰.۷۴۱۷
۵۰۰x۳	۰.۷۴۰۰	۰.۹۰۰۰	۰.۷۶۰۰	۰.۹۰۰۰	۰.۷۶۶۷	۰.۹۰۰۰	۰.۷۴۵۰	۰.۸۶۵۰	۰.۷۵۰۰	۰.۸۷۰۰
۱۰۰۰x۳	۰.۸۷۰۰	۰.۹۳۰۰	۰.۸۳۵۰	۰.۹۰۰۰	۰.۸۵۰۰	۰.۸۹۶۷	۰.۸۱۲۵	۰.۸۷۰۰	۰.۸۳۵۰	۰.۸۸۰۰
۲۰۰۰x۳	۰.۸۵۵۰	۰.۹۰۰۰	۰.۸۶۵۰	۰.۸۹۷۵	۰.۸۶۵۰	۰.۸۸۳۳	۰.۸۷۳۸	۰.۹۰۳۸	۰.۸۴۶۷	۰.۸۸۵۰

جدول ۱۷-نتایج حاصل از اجرای الگوریتم ParzenWindows روی مجموعه ای داده ای_۲ با درصد تست های مختلف و H=۰.۵ و H=۱ روی مجموعه های تست و train یکسان

	۰.۰۵ H=۱	۰.۰۵ H=۲	۰.۱ H=۱	۰.۱ H=۲	۰.۱۵ H=۱	۰.۱۵ H=۲	۰.۲ H=۱	۰.۲ H=۲	۰.۳ H=۱	۰.۳ H=۲
۵۴۰۴x۶	۰.۸۱۹۲	۰.۷۶۰۱	۰.۷۸۹۳	۰.۷۷۰۸	۰.۸۰۷۶	۰.۷۲۸۷	۰.۸۰۷۶	۰.۷۲۹۰	۰.۸۱۲۶	۰.۷۴۶۶

جدول ۱۸-نتایج حاصل از اجرای الگوریتم ParzenWindows روی مجموعه ای داده ای_phoneme با درصد تست های مختلف و H=۱ و H=۲ روی مجموعه های تست و train یکسان

	۰.۰۵ H=۰.۸	۰.۰۵ H=۱.۷	۰.۱ H=۰.۸	۰.۱ H=۱.۷	۰.۱۵ H=۰.۸	۰.۱۵ H=۱.۷	۰.۲ H=۰.۸	۰.۲ H=۱.۷	۰.۳ H=۰.۸	۰.۳ H=۱.۷
۱۵۰x۵	۱	۰.۷۵۰۰	۰.۸۶۶۷	۰.۹۳۳۳	۱	۰.۹۵۶۵	۱	۰.۹۰۰۰	۰.۹۱۱۱	۰.۸۶۶۷

جدول ۱۹-نتایج حاصل از اجرای الگوریتم ParzenWindows روی مجموعه ای داده ای_iris با درصد تست های مختلف و H=۰.۸ و H=۱.۷ روی مجموعه های تست و train یکسان

	۰.۰۵ H=۲۰	۰.۰۵ H=۶۰	۰.۱ H=۲۰	۰.۱ H=۶۰	۰.۱۵ H=۲۰	۰.۱۵ H=۶۰	۰.۲ H=۲۰	۰.۲ H=۶۰	۰.۳ H=۲۰	۰.۳ H=۶۰
۶۴۳۵x۳۷	۰.۴۸۱۳	۰.۸۴۷۸	۰.۵۱۲۴	۰.۸۴۱۶	۰.۴۶۶۸	۰.۸۵۵۰	۰.۴۸۴۸	۰.۸۴۶۱	۰.۴۵۰۵	۰.۸۳۹۴

جدول ۲۰-نتایج حاصل از اجرای الگوریتم ParzenWindows روی مجموعه ای داده ای_satimage با درصد تست های مختلف و h=۲۰ و h=۶۰ روی مجموعه های تست و train یکسان

۷.تحلیل نتایج

از جداول ۱-۵ مشاهده می شود که الگوریتم OneNN دارای کارایی بالایی می باشد؛ اگرچه این الگوریتم برای داده هایی نویزی مناسب نمی باشد و دارای سربار محاسباتی بالایی است، زیرا به ازای هر نمونه تست فاصله ی آن نمونه را با همه ی نمونه های train محاصبه می کند .

اگر فضای کار d بعدی باشد، آنگاه فاصله از فرمول زیر به دست می آید:

یک نمونه تست : $x = (x۱, x۲, \dots, xd)$

یک نمونه train : $y = (y۱, y۲, \dots, yd)$

$$distance = \sqrt[۲]{(x۱ - y۱)^۲ + (x۲ - y۲)^۲ + \cdots + (xd - yd)^۲}$$

که این کار را به ازای هر عنصر تست با همه ی عناصر train انجام می دهد؛ پس اگر تعداد عناصر train و تست به ترتیب برابر با n و m باشد، آنگاه پیچیدگی این روش برابر با $\theta = (nm)$ است.

از جداول ۶-۱۰ مشاهده می شود که الگوریتم Bayes نسبت به الگوریتم از کارایی کمتری بر خوردار است. در این الگوریتم هم با افزایش داده های تست و همچنین با افزایش اندازه ی کل داده ها، کارایی الگوریتم کاهش می یابد.

از جداول ۱۱-۱۵ مشاهده می شود که الگوریتم KNN از کارایی مناسبی بر خوردار است. در این الگوریتم برای k=۵ بهتر از k=۱۰ جواب می دهد. این الگوریتم از هر دو الگوریتم OneNN و Bayes بهتر عمل می کند.

از جداول ۱۶-۲۰ مشاهده می شود که الگوریتم ParzenWindows از همه ی الگوریتم های دیگر بهتر عمل می کند گو اینکه کارایی این الگوریتم خیلی به مقدار H بستگی دارد. این الگوریتم در dataset_۱ برای H=۰.۵، در dataset_۲ برای H=۱، در dataset_phoneme برای H=۱، در dataset_iris برای H=۰.۸، و در dataset_satimage برای H=۶۰ خیلی بهتر ، جواب می دهد.

به طور کلی در بین این چهار الگوریتم، ParzenWinodws در شرایطی که مقدار H مناسب انتخاب شود از سایر الگوریتم ها کارایی بهتری دارد.

۶.منابع

۱. **http://sci۲s.ugr.es/keel/dataset.php?cod=۱۰۵**

۲. **http://archive.ics.uci.edu/ml/datasets/Iris**

۳.**http://www.dice.ucl.ac.be/neural-nets/Research/Projects/ELENA/databases/REAL/satimaget**