

دانشگاه صنعتي امیرکبیر  
 (پلی‌تکنیک تهران)

دانشكده مهندسی کامپیوتر

گزارش پروژه‌ي پاياني درس يادگيري ماشين

عنوان مقاله:

زيرنمونه‌برداري براي روش‌هاي تركيبي و بدون نظارت جهت كشف داده‌هاي پرت به صورت كارآمد و مؤثر

نام دانشجو: سید احمد نقوی نوزاد

ش-د: 94131060

استاد درس:

دکتر ناظرفرد

بهار 1395

****

## مقدمه:

در ابتدا تعریف مختصری از داده‌ی پرت ارائه داده و سپس به لزوم کشف داده‌های پرت پرداخته و در ادامه به دو دسته‌بندی کلی از انواع روش‌های کشف داده‌های پرت اشاره خواهیم نمود. در نهایت به بررسی مختصری در مورد رویکرد پیشنهادشده در این مقاله خواهیم پرداخت.

یک داده‌ی پرت به یک داده‌ی ضبط‌شده و یا مجموعه‌ای از داده‌ها اطلاق می‌گردد که بنا به ظاهر نسبت به سایر مجموعه داده ناسازگار بوده و رفتار غیر نرمالی از خود بروز می‌دهند. حال این رفتار می‌تواند نسبت به کلیت مجموعه داده مورد بررسی قرار گیرد (داده‌ي پرت سراسري يا Global Outlier) و یا هم در حیطه‌ی یک همسایگی خاص از داده‌ی مورد نظر (مانندK نزدیک‌ترین همسایه) مورد توجه واقع گردد (داده‌ي پرت محلي يا Local Outlier). در این مقاله ما به طور خاص به دنبال کشف داده‌های پرت محلی با استفاده از روش‌های اصطلاحاً ترکیبی (ensemble) خواهیم بود.

کشف داده‌های پرت از اهمیت ویژه‌ای در بسیاری از کاربردهای عملی، نظیر کشف خطاهای اندازه‌گیری توسط حسگرهای مختلف (که به نوعی با داده‌ی پرت به عنوان نویز رفتار می‌کنند)، کشف سوء استفاده از کارت‌های اعتباری و یا هم یافتن اندازه‌گیری‌های غیرمتعارف در داده‌های علمی دارد؛ چرا که رفتار غیرمتعارفی که داده‌های پرت نسبت به باقی داده‌ها از خود بروز می‌دهند، می‌تواند موجب بروز مشکلات عدیده گشته و در نتیجه نیازمند توجهات خاصی بوده و البته می‌توانند سبب به وجود آمدن بینش‌های جدیدی نیز نسبت به مجموعه‌داده‌ی مورد بررسی گردند.

روش‌ها و رویکردهای متعددی جهت کشف داده‌های پرت تاکنون ارائه شده‌اند، که البته هر کدام با توجه به نوع داده‌ی پرتی که بر روی آن متمرکز شده‌اند و یا هم کاربرد و زمینه‌ی خاصی که نیازمند کشف داده‌ی پرت می‌باشد، با دیگران تفاوت دارند. دسته‌ی اول، رویکردهای آماری پارامتریک[[1]](#footnote-1) می‌باشند، که سعی دارند تا با استفاده از تخمین پارامترهای یک تخمین فرضی، آن تخمین را به مجموعه‌داده‌ی مورد بررسی نسبت دهند. اما مشکل موجود در مورد این نوع رویکردهای مبتنی بر یک توزیع فرضی‌ِ خاص، این است که پارامترهای توزیع مدنظر نسبت به حضور داده‌های پرت حساس بوده و ممکن است در محاسبه‌ی آن‌ها دچار خطا شویم. از جمله اثرات مضری که داده‌های پرت بر روی تخمین پارامترهای توزیع مد نظر می‌گذارند، می‌توان به پدیده‌هایی تحت عنوان «پوشش»[[2]](#footnote-2) و نیز «غرقانیدن»[[3]](#footnote-3) اشاره نمود. و اما دسته‌ی دوم از رویکردهای مورد استفاده جهت کشف داده‌های پرت، با نام رویکردهای غیرپارامتریک[[4]](#footnote-4) شناخته می‌شوند، که بر عکس مورد قبلی، توزیع خاصی را به داده‌ها نسبت نداده بلکه سعی در آن دارند تا به صورت آشکارا و یا هم ضمنی، وجهه‌های خاصی از تابع توزیع چگالی احتمالاتی را تخمین بزنند. از جمله روش‌های مبتنی بر این نوع رویکرد، روش‌های «مبتنی بر فاصله» و روش‌های «مبتنی بر چگالی» می‌باشند که البته این روش‌ها نیز می‌توانند مانند مورد قبل، از اثرات مضری مشابه پوشش و غرقانیدن که در نتیجه‌ی عمل تخمین‌زدن فاصله و چگالی صورت می‌گیرد، آسیب ببینند.

و اما در مورد رویکرد مورد استفاده در این مقاله باید گفت که با توجه به این مطلب که تکنیک‌های ترکیبی در مورد کشف داده‌های پرت کمتر مورد مطالعه قرار گرفته‌اند، لذا قصد داریم تا به «زیرنمونه‌برداری» تحت عنوان یک تکنیک خاص بپردازیم، که سبب می‌شود تا تنوع میان انواع روش‌های کشف داده‌ی پرت آشکار گردد. هم‌چنین به صورت آماری و مبتنی بر آزمایشات نشان خواهیم داد که یک روش کشف‌کننده‌ی داده‌ی پرت، بر روی یک زیرنمونه به خودی خود، علاوه بر اینکه سبب استنتاج تنوع میان انواع روش‌ها می‌گردد، تحت شرایط خاصی از عملکرد همان روشِ کشف داده‌ی پرت بر روی کل مجموعه‌داده پیشی می‌گیرد. البته که ساخت یک روش ترکیبی که بر روی زیرنمونه‌های متعدد کار می‌کند، می‌تواند سبب بهبود چشمگیر نتایج گردد؛ و نیز ذکر این نکته ضروری می‌نماید که روش‌های ترکیبی مطرح‌شده تاکنون در مقالات متعدد، روش‌های بانظارت[[5]](#footnote-5) بوده و تلاش در دسته‌بندی[[6]](#footnote-6) داده‌ها دارند، ولی در این مقاله تلاش خواهیم داشت تا به صورت آماری، صحت عملکرد روش‌های ترکیبی بی‌نظارت[[7]](#footnote-7) و مبتنی بر خوشه‌بندی[[8]](#footnote-8) در کشف داده‌های پرت را به اثبات برسانیم. در پایان باید گفت که استفاده از ترکیبی از روش‌های کشف داده‌های پرت بر روی زیرنمونه‌های متعدد از مجموعه‌داده، با توجه به اندازه‌ی زیرنمونه‌ها[[9]](#footnote-9) و نیز حجم کلی روش‌های ترکیب‌شده[[10]](#footnote-10)، به طور کلی نسبت به استفاده از یک روش منفرد کشف داده‌ی پرت بر روی کل مجموعه‌داده عملکرد بهتری خواهد داشت.

## شرح روش و پارامترها:

در اين مقاله اين‌گونه نيست كه تنها از مجموعه‌داده نمونه‌برداري كرده و سپس الگوريتم كشف داده‌ي پرت را بر روي زيرنمونه‌هاي مربوطه و با احتساب حضور ساير داده‌هاي خارج از زيرنمونه اجرا نمائيم، چرا كه با اين كار اطلاعات زيادي در مورد ماهيت پرت‌بودن بسياري از داده‌ها از بين رفته و نيز بسياري از داده‌ها نيز تنها امتيازي حسب پرت‌بودن از تنها برخي از زيرنمونه‌ها خواهند برد. بلكه در عوض ما در اين مقاله، به ازاي هر عضو ensemble، يك زيرنمونه از مجموعه داده انتخاب كرده و همسايگي مربوطه را تنها به ازاي داده‌هاي موجود در زيرنمونه براي هر كدام از الگوريتم‌هاي كشف داده‌ي پرت محاسبه مي‌نمائيم، كه اين روش سبب افزايش سرعت قابل توجهي در مقايسه با ساير روش‌هاي ensemble گشته و نيز نتايج نهائي بهتري را حاصل مي‌نمايد.

در اين مقاله يك روش پايه را با نام Feature Bagging ، به عنوان رقيب روش ensemble معرفي‌شده در اين مقاله مورد استفاده قرار مي‌دهيم و براي هر دوي روش‌هاي تركيبي (ensemble و Feature Bagging) يك مقدار معين براي تعداد اعضاي ensemble يعني 25 را انتخاب خواهيم نمود. در مورد روش رقيب يعني FeatureBagging نيز جهت تركيب نتايج نهائي، از عمل ميانگين‌گيري استفاده خواهيم نمود. در مورد روش ensemble نيز از زيرنمونه‌هاي با اندازه‌هاي متفاوت جهت ارزيابي نتايج نهائي استفاده نموده و در مورد روش‌هاي پايه نيز از اندازه‌هاي مختلف k براي يافتن نزديك‌ترين همسايه‌ها استفاده مي‌كنيم. در نهايت نتايج نهائي را در قالب‌هاي زير نشان خواهيم داد: i) مقدار ثابت k و اندازه‌هاي مختلف زيرنمونه؛ ii) مقدار ثابت زيرنمونه و مقادير مختلف k؛ iii) مقادير ثابت براي هر دوي اندازه‌ي زيرنمونه و مقدار k. در ضمن مقداري را براي k انتخاب خواهيم نمود كه در مورد روش پايه به نتايج مطلوبي منجر گردد.

هم‌چنين براي ارزيابي نتايج نهائي از ROC AUC[[11]](#footnote-11) استفاده مي‌نمائيم كه مقادير مختلف True Positive Rate را بر حسب مقادير False Positive Rate رسم نموده و نيز يك معيار عمومي و پركاربرد جهت ارزيابي روش‌هاي مختلف كشف داده‌هاي پرت مي‌باشد.

## مجموعه داده‌هاي مورد استفاده

در اين‌جا دو مجموعه داده‌ي مستقل دست‌ساز و مصنوعي[[12]](#footnote-12) را توليد مي‌نمائيم (batch1 و batch2) كه هر كدام از متشكل از 30 مجموعه‌داده‌ي مختلف با ابعاد، تعداد خوشه‌ها و نيز اندازه‌ي متفاوت خوشه‌ها مي‌باشند. مجموعه‌ي داده‌هاي مربوطه را با استفاده از توزيع گاوسين مخلوط[[13]](#footnote-13) توليد مي‌نمائيم و براي اين كار از مقادير مختلف ميانگين و انحراف از معيار كه از هر كدام در بازه‌هاي مشخصي مي‌باشند استفاده مي‌نمائيم. در نهايت مي‌بايست مجموعه داده‌هاي مربوطه را در فضاي چندبعدي چرخش[[14]](#footnote-14) دهيم كه براي اين‌كار نيز ابتدا يك ماتريس A را با داده‌هاي رندوم نرمال توليد نموده و سپس با ضرب ترانهاده‌ي اين ماتريس در خود آن، يك ماتريس واريانس-كوواريانس به دست مي‌آوريم كه خاصيت لازم به اصلاح positive-semiDefinite در مورد آن برقرار مي‌باشد. بعد ازين با استفاده از تابع mvnrnd()، يك توزيع گاوسين مخلوط را در مورد داده‌هاي مربوط به هر كدام از خوشه‌هاي يك مجموعه‌داده‌ي مشخص به دست مي‌آوريم. جهت اين‌كه داده‌هاي پرت محلي را در مورد اين مجموعه‌داده‌ها برچسب‌گذاري نمائيم، در ابتدا فاصله‌ي ماهالانوبيس[[15]](#footnote-15) را در مورد داده‌هاي موجود درون هر خوشه از مركز خوشه محاسبه نموده و سپس آن داده‌هايي را كه اين فاصله براي آن‌ها بيشتر از چارك 0.975 توزيع مربوط به فاصله‌هاي داده‌هاي درون خوشه باشد، به عنوان داده‌ي پرت برچسب‌گذاري مي‌نمائيم.

در مورد مجموعه‌داده‌هاي واقعي مورد استفاده نيز، در اين‌جا از مجموعه‌داده‌‌هاي معروف Satimage، Segment (كه به سه مجموعه‌داده‌ي مجزا با داده‌هاي پرت متفاوت تبديل خواهد شد)، Wisconsin Breast Cancer (WBC) و نيز مجموعه‌داده‌ي Waveform Database Generator (waveform) استفاده خواهيم نمود. در مورد هر كدام از اين مجموعه‌داده‌ها در ابتدا آن كلاسي را كه تعداد داده‌هاي كمتري را داراست به عنوان طعمه انتخاب نموده و سپس مانند مجموعه‌داده‌هاي مصنوعي كه پيش ازين قيد گرديد، آن داده‌هايي را كه فاصله‌ي ماهالانوبيس آن‌ها از مركز خوشه، بيشتر از چارك 0.9 توزيع مربوط به فاصله‌هاي داده‌هاي درون خوشه باشد، به عنوان داده‌ي پرت برچسب‌گذاري مي‌نمائيم.

در نهايت از الگوريتم‌هاي كشف داده‌هاي پرت محلي جهت شناسائي داده‌هاي پرت در مجموعه‌داده‌هاي نامبرده استفاده خواهيم نمود.

## نتايج مربوط به پياده‌سازي

1. مجموعه‌داده‌هاي مصنوعي

در اين‌جا مجموعه‌داده‌ي مصنوعي batch1 متشكل از 30 مجموعه‌داده‌ي مجزا را به عنوان كانديد برگزيده و دو روش پايه جهت كشف داده‌هاي پرت محلي با نام‌هاي LOF[[16]](#footnote-16) و LoOP[[17]](#footnote-17) را بر روي آن‌ها پياده نموده و روش‌هاي تركيبي قيدشده پيش ازين، با نام‌هاي FeatureBagging و ensemble را نيز با پارامترهاي مشخص (اندازه‌ي زيرنمونه‌هاي مختلف و مقدار مشخص k=3) بر روي آن‌ها پياده مي‌نمائيم. در اين‌جا مقادير ROC AUC مختلف به ازاي زيرنمونه‌هاي مختلف را به دست آورده و براي رسم آن‌ها از تابع boxplot() استفاده مي‌نمائيم. نتايج بيان‌شده در مقاله و نيز نتايج حاصله از پياده‌سازي در ادامه مي‌آيد:

|  |  |
| --- | --- |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي مصنوعي batch1 و روش پايه‌ي LOF و k=3 |
|  | نتايج بيان‌شده در مقاله در مورد مجموعه‌داده‌ي مصنوعي batch1 و روش پايه‌ي LOF و k=3 |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي مصنوعي batch1 و روش پايه‌ي LoOP و k=3 |
|  | نتايج بيان‌شده در مقاله در مورد مجموعه‌داده‌ي مصنوعي batch1 و روش پايه‌ي LoOP و k=3 |
| شكل 1. نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي مصنوعي batch1 و روش‌هاي پايه‌ي مختلف | |

متأسفانه همانطور كه قابل مشاهده است، نتايج حاصل از پياده‌سازي با نتايج قيد‌شده در مقاله يكسان نمي‌باشد و علت اين مسئله را نيز مي‌توان در تصادفي‌بودن رويه‌ي توليد داده‌هاي مصنوعي جستجو نمود.

با توجه به نكات قيدشده درمقاله انتظار آن بود كه با افزايش اندازه‌ي زيرنمونه، شاهد كاهش محدوده‌ي توزيع مقادير ROC AUC باشيم كه در اين‌جا اين‌گونه نشد، اما در مورد برخي مجموعه‌داده‌هاي واقعي كه در ادامه خواهد آمد، نتايج خوبي را مطابق مندرجات موجود در مقاله شاهد خواهيم بود.

1. مجموعه‌داده‌هاي واقعي

در ابتدا به مجموعه‌داده‌هاي WBC و Satimage خواهيم پرداخت كه تنها در مورد دومي نتايج مطلوبي را شاهد خواهيم بود. نتايج مربوط به مجموعه‌داده‌ي WBC به شرح ذيل مي‌باشند:

|  |  |
| --- | --- |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي WBC و روش پايه‌ي LOF و k=2 |
|  | نتايج بيان‌شده در مقاله در مورد مجموعه‌داده‌ي WBC و روش پايه‌ي LOF و k=2 |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي WBC و روش پايه‌ي LoOP و k=2 |
|  | نتايج بيان‌شده در مقاله در مورد مجموعه‌داده‌ي WBC و روش پايه‌ي LoOP و k=2 |
| شكل 2. نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي WBC و روش‌هاي پايه‌ي مختلف | |

در ادامه نتايج مربوط به مجموعه‌داده‌ي Satimage ذكر مي‌گردد. لازم به ذكر است كه با توجه به بزرگ‌تر بودن مجموعه‌داده‌ي Satimage ، مقدار k را نيز بزرگ‌تر و برابر 50 انتخاب مي‌نمائيم. در ادامه داريم:

|  |  |
| --- | --- |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي Satimage و روش پايه‌ي LOF و k=50 |
|  | نتايج بيان‌شده در مقاله در مورد مجموعه‌داده‌ي Satimage و روش پايه‌ي LOF و k=50 |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي Satimage و روش پايه‌ي LoOP و k=50 |
|  | نتايج بيان‌شده در مقاله در مورد مجموعه‌داده‌ي Satimage و روش پايه‌ي LoOP و k=50 |
| شكل 3. نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي Satimage و روش‌هاي پايه‌ي مختلف | |

در اين قسمت، اندازه‌ي زيرنمونه را ثابت و برابر 0.1 در نظر گرفته و نتايج را بر حسب مقادير متفاوت k بر روي مجموعه داده‌ي Waveform و براي روش‌هاي پايه، تركيب زيرنمونه‌ها[[18]](#footnote-18) و تركيب چينش‌هاي مختلفي از ويژگي‌ها[[19]](#footnote-19) نمايش مي‌دهيم. در ادامه داريم:

|  |  |
| --- | --- |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي Waveform و روش پايه‌ي LOF و sampleSize=0.1 |
|  | نتايج بيان‌شده در مقاله در مورد مجموعه‌داده‌ي Satimage و روش پايه‌ي LOF و sampleSize =0.1 |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي Waveform و روش پايه‌ي LoOP و sampleSize=0.1 |
|  | نتايج بيان‌شده در مقاله در مورد مجموعه‌داده‌ي Waveform و روش پايه‌ي LoOP و sampleSize=0.1 |
| شكل 4. نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌ي Waveform و روش‌هاي پايه‌ي مختلف | |

علاوه بر اينكه نتايج حاصل از پياده‌سازي خوشبختانه با نتايج مندرج در مقاله همخواني بالائي دارند! بايد گفت كه با افزايش مقدار k شاهد افزايش اندك اما پايدار در مقدار ROC AUC براي روش‌هاي پايه (LOF و LOoP) و Subsampling Ensemble خواهيم بود و به عبارتي هر دوي اين روش‌ها از يك الگوي يكسان و پايدار پيروي مي‌نمايند، در حالي كه براي روش FeatureBagging شاهد نوسان و واريانس بالاتري در مورد مقادير ROC AUC هستيم.

در ادامه نتايج حاصله در مورد مجموعه‌داده‌ي Segment را با مقادير مشخص براي پارامترها (k=20 كه در مورد اين مجموعه‌داده و البته روش‌هاي پايه‌ي مختلف نتايج خوبي را حاصل نموده است؛ و subsampleSize=0.1) به صورتي كه در ادامه مي‌آيد نشان مي‌دهيم:

|  |  |
| --- | --- |
|  | نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌هاي Segment و روش‌هاي پايه‌ي LOF(1) و LoOP(2) و k=20 وsampleSize=0.1 |
|  | نتايج مندرج در مقاله در مورد مجموعه‌داده‌هاي Segment و روش‌هاي پايه‌ي LDOF، LOF و LoOP و k=20 وsampleSize=0.1 |
| شكل 5. نتايج حاصل از پياده‌سازي بر روي مجموعه‌داده‌هاي Segment و روش‌هاي پايه‌ي مختلف | |

همانطور كه قابل مشاهده است، روش زيرنمونه‌برداري معرفي‌شده در اين مقاله به مراتب نتايج بهتري را نسبت به روش‌هاي پايه و البته روش رقيب FeatureBagging حاصل نموده است، كه اين خود گواه بر صحت ادعاهاي قيدشده در مقاله مي‌باشد.

## معرفي ايده‌ي پيشنهادي و نتايج پياده‌سازي

تا اين‌جا ما از روش‌هاي تركيبي، جهت كشف داده‌هاي پرت محلي استفاده نموديم؛ حال قصد داريم تا روشي را جهت كشف داده‌هاي پرت سراسري معرفي نمائيم. به اين ترتيب كه ابتدا يك مجموعه‌داده‌ي مصنوعي را مانند آن‌چه پيش از اين قيد گرديد تهيه مي‌نمائيم، تنها با اين تفاوت كه تعدادي داده‌ي جديد را با مقادير بسيار دورتر از ميانگين هر بعد به مجموعه داده اضافه نموده و آن‌ها را تحت عنوان داده‌ي پرت سراسري برچسب مي‌زنيم. حال در عمل پياده‌سازي ابتدا مجموعه‌داده را با استفاده از تابع kmeans() و مقادير متفاوت clustNo (كه همان تعداد خوشه‌هاست) خوشه‌بندي نموده و سپس براي هر داده يك مقدار احتمالاتي بين 0 و 1 را به عنوان ضريب داده‌ي پرت سراسري معرفي مي‌نمائيم، كه اين مقدار از تقسيم فاصله‌ي هر داده از مركز خوشه‌اي كه در آن قرار دارد بر بيشينه‌ي فاصله‌‌هاي داده‌هاي موجود درون خوشه از مركز به دست مي‌آيد و البته عددي مابين 0 و 1 مي‌باشد. در نهايت به ازاي خوشه‌بندي‌هاي مختلف، مقادير متفاوتي براي اين ضرايب خواهيم داشت. سپس جهت به دست‌آوردن مقدار نهائي براي ضريب مربوطه به ازاي هر داده، از روش اول‌عمق استفاده مي‌نمائيم، كه در آن ابتدا ماتريس متشكل از ضرايب (سطرها بيانگر داده‌ها و ستون‌ها بيانگر تعداد خوشه‌ها) را به صورت ستوني و هر ستون مجزا از ديگري، در قالب ترتيب نزولي مرتب مي‌نمائيم. سپس از سطر اول شروع نموده و سطر به سطر جلو مي‌رويم و اولين مقدار ضريب كه به ازاي هر داده مشاهده مي‌كنيم را به عنوان ضريب مربوط به آن داده گزارش مي‌نمائيم تا زماني كه به تمامي داده‌ها مقدار ضريب مربوطه نسبت داده شود. به عبارتي در اين روش ضريبي براي هر داده از ماتريس مرتب‌شده صورت نزولي انتخاب مي‌شود، كه در ميان خوشه‌بندي‌هاي مختلف، در بالاترين رديف ممكن قرار دارد.

در نهايت، بردار ضرايب داده‌ي پرت سراسري را به همراه بردار مربوط به برچسب‌ها به تابع perfcurve() مي‌دهيم تا منحني ROC مربوط به روش پيشنهادي را براي ما رسم نموده و مساحت زير نمودار (AUC) را نيز براي ما فراهم نمايد. هر چه اين مساحت بيشتر باشد، عملكرد الگوريتم پيشنهادي بهتر بوده است. لازم به ذكر است كه در اين‌جا ما در عمل خوشه‌بندي با استفاده از kmeans()، از دو معيار فاصله‌ي SquaredEuclidean Distance و نيز Cosine Distance استفاده مي‌نمائيم كه البته در مورد دومي به نتايج بهتري دست پيدا مي‌كنيم. در ادامه داريم:

|  |  |
| --- | --- |
|  | منحني ROC براي روش پيشنهادي، با معيار فاصله‌ي SquaredEuqlidean Distance براي الگوريتم kmeans  AUC = 0.8648 (x) |
|  | منحني ROC براي روش پيشنهادي، با معيار فاصله‌ي Cosine Distance براي الگوريتم kmeans  AUC = 0.9941 (√) |
| شكل 6. نتايج حاصل از پياده‌سازي روش پيشنهادي جهت كشف داده‌ي پرت سراسري | |

## مراجع

|  |  |
| --- | --- |
| **Zimek, Arthur, et al. "Subsampling for efficient and effective unsupervised outlier detection ensembles." *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013.** | [1] |
| **Breunig, Markus M., et al. "LOF: identifying density-based local outliers." *ACM sigmod record*. Vol. 29. No. 2. ACM, 2000.** | [2] |
| **Kriegel, Hans-Peter, et al. "LoOP: local outlier probabilities." *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009.** | [3] |

1. Parametric Statistical Approaches [↑](#footnote-ref-1)
2. Masking [↑](#footnote-ref-2)
3. Swamping [↑](#footnote-ref-3)
4. Non-parametric Approaches [↑](#footnote-ref-4)
5. Supervised [↑](#footnote-ref-5)
6. Classification [↑](#footnote-ref-6)
7. Unsupervised [↑](#footnote-ref-7)
8. Clustering [↑](#footnote-ref-8)
9. Subsample size [↑](#footnote-ref-9)
10. Ensemble size [↑](#footnote-ref-10)
11. Area under the receiver operating characteristic curve (ROC AUC) [↑](#footnote-ref-11)
12. Synthetic [↑](#footnote-ref-12)
13. Gaussian Mixture Distributions [↑](#footnote-ref-13)
14. Rotation [↑](#footnote-ref-14)
15. Mahalanobis [↑](#footnote-ref-15)
16. Local Outlier Factor (LOF) [↑](#footnote-ref-16)
17. Local Outlier Probability (LoOP) [↑](#footnote-ref-17)
18. Subsampling Ensemble [↑](#footnote-ref-18)
19. Feature Bagging Ensemble [↑](#footnote-ref-19)