# شناسایی خطاهای داده‌های خام بویه‌های موج‌نگار با استفاده از روش ضریب داده پرت محلی

کیومرث محمودی[۲]، محمد جواد کتابداری[۱]

ketabdar@aut.ac.ir

۱- کارشناس ارشد، دانشکده مهندسی دریا، دانشگاه صنعتی امیرکبیر
۲- دانشیار، دانشکده مهندسی دریا، دانشگاه صنعتی امیرکبیر

**چکیده**

استخراج مشخصات دریا معمولاً از طریق بویه‌های موج نگار انجام می‌شود. اما ثبت داده توسط موج نگارها معمولاً با خطاهایی همراه است. لذا قبل از استخراج هرگونه اطلاعاتی لازم است این خطاها را شناخت و آنها را حذف و یا تصحیح کرد. هدف از این تحقیق، شناسایی خطاهای موجود در برداشت داده‌های خام از بویه‌های موج نگار، با استفاده از روش ضریب داده پرت محلی (LOF) است. LOF روشی قدرتمند جهت شناسایی ناهنجاری داده‌ها در یادگیری ماشین است، که در بسیاری از کاربردهای عملی از آن استفاده می‌شود. در این مقاله داده‌های روزانه بویه موج‌نگار سازمان بنادر و دریانوردی که در آب‌های خلیج فارس در منطقه عسلویه استان بوشهر به آب انداخته شده است در بازه ۹۲/۲/۱۷ تا ۹۲/۵/۱ مورد تحلیل خطا به این روش قرار گرفت. نتایج نشان می‌دهد LOF روشی کارآمد برای شناسایی خطاهای موجود در داده‌های موج نگار است.

**واژگان کلیدی: موج نگار، داده موج، خلیج فارس، شناسایی خطا، روش ضریب داده پرت محلی (LOF).**

| | |
|---|---|
| تاریخ دریافت مقاله : | ۹۳/۱۲/۱۸ |
| تاریخ پذیرش مقاله : | ۹۵/۰۱/۱۶ |

**۱ - مقدمه**

داده‌های حاصل از یک مشاهده وقتی مفید و کارآمد است که نماینده واقعی موضوع مورد مطالعه باشد. برای بدست آوردن نتایج مناسب و مطلوب از داده‌های جمع آوری شده، داده‌ها باید نماینده واقعی مشاهدات بوده، و به نحو مفیدی تهیه و تنظیم شده باشند تا قابل تجزیه و تحلیل بوده و بتوان به کمک آنها به نتیجه‌گیری صحیح رسید. اهمیت برداشت داده‌های مناسب و با کمترین خطا در مطالعات آزمایشگاهی از آنجا دارای اهمیت است که معمولاً از این داده‌ها، در صحت سنجی و کالیبراسیون مدل‌های عددی و ریاضی استفاده می‌شود [۱].

داده‌های خطا نتایج تجزیه و تحلیل داده‌ها را تحت تاثیر قرار داده و منجر به اختلال در نتیجه گیری از اطلاعات خواهند شد. داده‌های خطا می‌توانند بر اثر عوامل مختلفی به وجود آیند. دستگاه‌های فیزیکی برای تخمین اندازه‌گیری‌ها ممکن است با یک سری عوامل گذرا که موجب بد عمل کردن آنها می‌شود مواجه باشند. در نتیجه ممکن است خطایی در انتقال اطلاعات یا استنتاج نهایی بوجود آید. همچنین این داده‌ها ممکن است بر اثر تغییر در رفتار سیستم، خطاهای انسانی، خطاهای دستگاهی یا به طور ساده به خاطر انحرافهای طبیعی در جمعیت نمونه‌ها به وجود آیند. تعریف ریاضیاتی دقیقی برای تشخیص منبع داده خطا وجود ندارد. تشخیص اینکه یک مشاهده خطا است یا نه، صرفاً یک فرایند کیفی و ابتکاری است که به تجربه شخصی فرد و کاربردی که از داده‌ها داریم وابسته است.

برای اندازه‌گیری مشخصه‌های دریایی از دستگاه‌های متنوعی استفاده می‌شود. این دستگاه‌ها قادر به جمع‌آوری اطلاعات هواشناسی و اقیانوس شناسی مانند سرعت، جهت و مدت وزش باد، دما، فشار، رطوبت هوا، غلظت و شوری آب دریا و تاریخچه زمانی تغییرات تراز آب هستند[۲]. بویه‌های موج نگار و پلات فرم‌ها از وسایل متداول این اندازه‌گیری‌ها هستند. هر یک از این دستگاه‌ها می‌توانند اطلاعات مهم موج از جمله ارتفاع و پریود آن را از داده‌های خام استخراج کنند. ثبت داده توسط موج نگارها معمولاً با خطاهای مختلفی همراه است که قبل از استخراج هرگونه اطلاعاتی لازم است این خطاها را شناخت و آن‌ها را حذف و یا تصحیح نمود. هدف از این تحقیق، شناسایی داده‌های مشکوک به خطا موجود در داده‌های برداشت شده از بویه‌ها با استفاده از روش ضریب داده پرت محلی[1] (LOF) است. در این مطالعه، داده‌ای به عنوان کاندیدای مشکوک به خطا در نظر گرفته شده است که به طور قابل توجهی متناقض با سایر اعضای نمونه ظاهر شده، و احتمالا با یک فرآیند متفاوت تولید شده است.

**۲- خطاهای احتمالی در رکورد موج**

در این بخش، برخی از خطاهای احتمالی موجود در برداشت داده‌های خام از بویه‌های موج‌نگار معرفی شده است:

- وجود Trend در موج: گاهی تراز متوسط داده‌های زمانی ثبت شده توسط بویه‌ها شیبی به سمت بالا یا پایین دارد. عموماً این مسئله بواسطه اختلال در دستگاه‌ها و تجهیزات است. گاهی ممکن است ترند به دلیل تأثیر تغییرات جزر و مد بر روی سنسور فشار بویه بروز نماید.
- وجود Noise در موج: نویز به وجود پارازیت و ارتعاشات با فرکانس بالا در تاریخچه زمانی موج می‌گویند. نویز ممکن است به دلیل حرکت پروانه شناورهای عبوری، یا تشدیدهای موردی در سیستم موج نگار و مهار آن ایجاد شود.
- وجود MWL در موج: معمولاً در اثر عبور موج به طور طبیعی تراز سطح آب بالا[2] یا پایین[3] می‌رود. گاهی نیز این تغییرات تراز سطح آب در صورتی که شدید باشد به علت وجود اختلال Drift در دستگاه است[۲].

**۳- مواد و روش‌ها**

داده‌های مورد استفاده در این تحقیق، داده‌های گردآوری شده از یکی از بویه‌های موج‌نگار موجود در آب‌های خلیج فارس است که در عسلویه استان بوشهر به آب انداخته شده است. این بویه در طول جغرافیایی ۲۵،۵ و عرض جغرافیایی ۵۲،۵ واقع شده است. بویه مورد استفاده از نوع Wavescan است (شکل (۱)). این بویه یک ابزار چند کاره و یکی از پرکاربردترین بویه‌ها است در سرتاسر دنیا است.

[1] Local Outlier Factor (LOF)
[2] Set up
[3] Set down


**شکل (۱) بویه موج نگار Wavescan عسلویه.**

. از این دستگاه برای اندازه‌گیری و پایش داده‌های مختلف اقیانوس‌شناسی، هواشناسی و کیفیت آب استفاده می‌شود. در ادامه این بخش مشخصات این بویه به صورت مختصر معرفی شده است.

ویژگی‌ها:

- طراحی منحصر به فرد برای بهینه‌سازی اندازه‌گیری‌های جهت موج.
- پردازش داده‌های اندازه‌گیری شده.
- ارتباط دو طرفه برای برای انتقال و کنترل داده‌ها.
- انتقال و ارائه آنی داده‌های اندازه‌گیری شده.
- پیکربندی انعطاف‌پذیر سنسورها و جمع آوری داده‌ها.
- طراحی شده برای حمل و نقل و استقرار امن و آسان.
- حمل و نقل آسان و مونتاژ محلی.
- ارائه لینک‌های صوتی آبی به رصدخانه اعماق اقیانوس.
- طراحی ویژه به منظور به حداقل رساندن اثرات پهلوگیری و اغتشاش شناورها روی حرکت بویه.

کاربردها:

- طراحی و عملیات دریایی.
- مطالعات اقلیمی و هواشناسی.
- مناسب برای عملیات در آبهای عمیق.
- نظارت بر بندرگاه.
- مهندسی سواحل.
- مطالعات علمی.
- مطالعات انرژی موج.
- کنترل ترافیک دریایی.
- مطالعات نظارت بر کیفیت آب.

اندازه‌گیری جهت موج: این بویه مجهز به یک حسگر برای اندازه‌گیری‌های جهت موج مبتنی بر اصل اندازه‌گیری شیب است.

ارائه داده‌ها:

وضعیت پارامترهای سنسور انتخابی با استفاده از نرم‌افزار مخصوص به طور مداوم پایش شده بنابراین در صورت بروز هر گونه حادثه پیش‌بینی نشده‌ای یک پیام به آدرس از پیش تعریف شده ارسال می‌شود. گزارش‌های بویه به راحتی تولید و پیکربندی شده و برای خروجی به فرمت‌های مختلف آماده می‌شوند. این ویژگی امکان ارائه آنی داده‌ها روی اینترنت را فراهم می‌کند. همچنین نرم‌افزار مورد استفاده دارای روش‌هایی برای بررسی کیفیت داده‌ها است.

در این تحقیق داده‌های بویه از سایت سازمان بنادر و دریانوردی استخراج شده است [۳]. در جدول (۱)، داده‌های مورد بررسی به همراه فهرست مشخصه‌های اندازه‌گیری شده، آمده است. در این جدول، مقادیر مشخصه‌های ارتفاع موج شاخص، میزان رطوبت، فشار هوا، دمای آب و سرعت باد قرار دارد. هدف، شناسایی داده‌های مشکوک به خطا موجود در این اندازه‌گیری‌ها است.

### ۳-۱- روش ضریب داده پرت محلی

شناسایی داده‌های پرت (داده‌های مشکوک به خطا)، شاخه‌ای کاربردی و بسیار مهم از داده کاوی است که به شناسایی الگوهای متناقض با جمعیت نرمال نمونه، در یک جامعه آماری می‌پردازد. به علت اهمیت شناسایی داده‌های پرت در علوم مختلف، این موضوع به یکی از مهمترین موضوعات تحقیقاتی در آمار، یادگیری ماشین و داده‌کاوی تبدیل شده است. برخی از کاربردهای الگوریتم‌های شناسایی داده‌های پرت به صورت نمونه در زیر معرفی شده است:

- شناسایی تقلب: به عنوان مثال شناسایی تقلب در کارت‌های اعتباری و یا گوشی‌های تلفن همراه [۴].
- فرایندهای وام و وام دهی: شناسایی کلاهبرداری‌ها در امور وام و همچنین شناسایی مشتری‌هایی که پتانسیل ایجاد مشکل را دارند [۴].
- شناسایی نفوذ: به عنوان مثال شناسایی دسترسی‌های غیرمجاز به شبکه‌های کامپیوتری [۵].

**جدول (۱) داده‌های گردآوری شده از بویه موج‌نگار عسلویه.**

| شماره | تاریخ برداشت | ارتفاع موج شاخص (m) | رطوبت (درصد) | فشار هوا (اتمسفر) | دمای آب ($^{\circ}C$) | سرعت باد (m/sec) |
|---|---|---|---|---|---|---|
| ۱ | ۱۳۹۲/۲/۱۷ | ۰,۳۸ | ۶۲,۲۴ | ۱۰۰۴,۲۳ | ۲۸,۱۳ | ۱,۶۳ |
| ۲ | ۱۳۹۲/۲/۱۸ | ۰,۲۳ | ۵۲,۹۳ | ۱۰۰۵,۴۷ | ۲۸,۶۳ | ۱,۸۹ |
| ۳ | ۱۳۹۲/۲/۱۹ | ۰,۲۲ | ۶۹,۵۹ | ۱۰۰۵,۴۹ | ۲۵,۷۰ | ۰,۸۴ |
| ۴ | ۱۳۹۲/۲/۲۰ | ۰,۲۷ | ۷۲,۰۹ | ۱۰۰۴,۹۱ | ۲۵,۸۸ | ۱,۰۹ |
| ۵ | ۱۳۹۲/۲/۲۱ | ۰,۲۲ | ۷۴,۰۲ | ۱۰۰۳,۱۴ | ۲۶,۱۶ | ۱,۰۵ |
| ۶ | ۱۳۹۲/۲/۲۲ | ۰,۱۹ | ۷۱,۶۷ | ۱۰۰۲,۷۳ | ۲۶,۹۲ | ۱,۰۶ |
| ۷ | ۱۳۹۲/۲/۲۳ | ۰,۱۶ | ۶۸,۶۸ | ۱۰۰۲,۷۰ | ۲۷,۶۵ | ۰,۶۵ |
| ۸ | ۱۳۹۲/۲/۲۴ | ۰,۱۷ | ۶۸,۰۵ | ۱۰۰۲,۸۰ | ۲۸,۴۲ | ۱,۰۴ |
| ۹ | ۱۳۹۲/۲/۳۱ | ۰,۱۸ | ۷۷,۴۲ | ۱۰۰۳,۷۸ | ۲۷,۷۴ | ۱,۱۵ |
| ۱۰ | ۱۳۹۲/۳/۱ | ۰,۵۵ | ۵۳,۴۴ | ۱۰۰۴,۰۵ | ۲۸,۵۹ | ۱,۰۵ |
| ۱۱ | ۱۳۹۲/۳/۲ | ۰,۳۷ | ۶۰,۹۰ | ۱۰۰۳,۱۹ | ۲۸,۱۷ | ۱,۲۲ |
| ۱۲ | ۱۳۹۲/۳/۳ | ۰,۳۷ | ۶۳,۶۰ | ۹۹۸,۸۲ | ۲۷,۱۹ | ۱,۱۵ |
| ۱۳ | ۱۳۹۲/۳/۴ | ۰,۵۶ | ۷۶,۵۹ | ۹۹۹,۹۷ | ۲۷,۷۵ | ۲,۰۲ |
| ۱۴ | ۱۳۹۲/۳/۵ | ۰,۷۸ | ۸۴,۴۳ | ۱۰۰۱,۲۹ | ۲۷,۸۶ | ۲,۵۱ |
| ۱۵ | ۱۳۹۲/۳/۶ | ۰,۶۵ | ۷۹,۷۰ | ۱۰۰۱,۵۳ | ۲۸,۶۸ | ۱,۸۱ |
| ۱۶ | ۱۳۹۲/۳/۷ | ۰,۵۸ | ۸۱,۷۷ | ۱۰۰۱,۷۹ | ۲۹,۰۷ | ۱,۶۶ |
| ۱۷ | ۱۳۹۲/۳/۸ | ۱,۳۱ | ۷۴,۹۴ | ۱۰۰۱,۰۸ | ۲۹,۴۱ | ۲,۸۱ |
| ۱۸ | ۱۳۹۲/۳/۹ | ۲,۱۳ | ۵۶,۱۸ | ۱۰۰۰,۰۰ | ۲۸,۸۵ | ۴,۸۸ |
| ۱۹ | ۱۳۹۲/۳/۱۰ | ۲,۱۵ | ۶۶,۹۲ | ۹۹۸,۷۳ | ۲۸,۱۴ | ۳,۱۷ |
| ۲۰ | ۱۳۹۲/۳/۱۱ | ۱,۱۱ | ۷۶,۹۲ | ۹۹۸,۰۸ | ۲۸,۵۱ | ۲,۶۷ |
| ۲۱ | ۱۳۹۲/۳/۱۲ | ۱,۱۵ | ۷۹,۱۶ | ۹۹۶,۶۹ | ۲۹,۵۲ | ۱,۹۷ |
| ۲۲ | ۱۳۹۲/۳/۱۳ | ۱,۶۴ | ۷۹,۰۵ | ۹۹۷,۰۴ | ۲۹,۰۳ | ۳,۷۲ |
| ۲۳ | ۱۳۹۲/۳/۱۴ | ۰,۶۲ | ۷۶,۴۰ | ۱۰۰۰,۱۴ | ۲۹,۷۲ | ۲,۲۵ |
| ۲۴ | ۱۳۹۲/۳/۱۵ | ۰,۶۷ | ۷۵,۵۳ | ۱۰۰۰,۸۴ | ۳۰,۰۸ | ۱,۷۴ |
| ۲۵ | ۱۳۹۲/۳/۱۶ | ۱,۰۷ | ۷۶,۴۶ | ۱۰۰۰,۳۷ | ۳۰,۶۴ | ۲,۴۹ |
| ۲۶ | ۱۳۹۲/۳/۱۷ | ۱,۶۶ | ۵۷,۳۵ | ۹۹۶,۰۸ | ۳۰,۳۴ | ۵,۲۸ |
| ۲۷ | ۱۳۹۲/۳/۱۸ | ۲,۰۴ | ۶۲,۲۸ | ۹۹۴,۱۸ | ۲۹,۲۶ | ۶,۴۷ |
| ۲۸ | ۱۳۹۲/۳/۱۹ | ۲,۰۰ | ۵۵,۹۷ | ۹۹۴,۳۵ | ۲۸,۵۵ | ۴,۹۱ |
| ۲۹ | ۱۳۹۲/۳/۲۰ | ۱,۵۰ | ۷۶,۹۹ | ۹۹۴,۲۷ | ۲۸,۷۸ | ۱,۴۹ |
| ۳۰ | ۱۳۹۲/۳/۲۱ | ۱,۵۹ | ۷۸,۴۹ | ۹۹۴,۱۴ | ۲۸,۸۱ | ۱,۸۴ |
| ۳۱ | ۱۳۹۲/۳/۲۲ | ۲,۰۱ | ۶۸,۹۳ | ۹۹۲,۷۳ | ۲۸,۸۴ | ۳,۵۷ |
| ۳۲ | ۱۳۹۲/۳/۲۳ | ۱,۶۵ | ۷۹,۶۵ | ۹۹۳,۷۱ | ۲۸,۷۹ | ۳,۱۴ |
| ۳۳ | ۱۳۹۲/۳/۲۴ | ۰,۹۲ | ۸۰,۷۱ | ۹۹۴,۳۱ | ۲۹,۴۶ | ۲,۴۱ |
| ۳۴ | ۱۳۹۲/۳/۲۵ | ۰,۶۹ | ۷۷,۵۶ | ۹۹۳,۸۸ | ۳۰,۱۰ | ۱,۸۴ |
| ۳۵ | ۱۳۹۲/۳/۲۶ | ۱,۷۳ | ۸۲,۰۵ | ۹۹۲,۳۳ | ۳۰,۳۹ | ۲,۹۲ |
| ۳۶ | ۱۳۹۲/۳/۲۷ | ۲,۰۷ | ۷۵,۴۰ | ۹۹۱,۴۲ | ۳۰,۶۳ | ۳,۲۳ |
| ۳۷ | ۱۳۹۲/۳/۲۸ | ۱,۸۹ | ۷۱,۹۹ | ۹۹۱,۳۷ | ۳۱,۰۰ | ۳,۲۸ |
| ۳۸ | ۱۳۹۲/۳/۲۹ | ۲,۱۵ | ۵۲,۹۶ | ۹۹۳,۳۹ | ۳۰,۸۳ | ۶,۲۰ |
| ۳۹ | ۱۳۹۲/۳/۳۰ | ۲,۱۷ | ۶۹,۱۲ | ۹۹۶,۷۷ | ۳۰,۴۱ | ۳,۳۶ |
| ۴۰ | ۱۳۹۲/۳/۳۱ | ۱,۸۹ | ۷۰,۶۹ | ۹۹۶,۷۰ | ۳۰,۶۵ | ۳,۱۴ |
| ۴۱ | ۱۳۹۲/۴/۱ | ۱,۵۳ | ۷۶,۴۳ | ۹۹۶,۲۶ | ۳۰,۸۱ | ۳,۶۷ |

| | | | | | | |
|---|---|---|---|---|---|---|
| ۴۲ | ۱۳۹۲/۴/۲ | ۰,۷۵ | ۷۸,۶۲ | ۹۹۵,۹۸ | ۳۱,۳۴ | ۲,۱۰ |
| ۴۳ | ۱۳۹۲/۴/۳ | ۰,۷۹ | ۷۹,۶۰ | ۹۹۴,۴۵ | ۳۱,۸۵ | ۱,۸۲ |
| ۴۴ | ۱۳۹۲/۴/۴ | ۰,۹۵ | ۷۸,۷۶ | ۹۹۳,۳۹ | ۳۱,۷۷ | ۲,۵۷ |
| ۴۵ | ۱۳۹۲/۴/۵ | ۰,۸۷ | ۸۰,۶۵ | ۹۹۵,۱۰ | ۳۲,۲۰ | ۱,۷۹ |
| ۴۶ | ۱۳۹۲/۴/۶ | ۱,۱۱ | ۷۷,۲۷ | ۹۹۴,۷۷ | ۳۲,۲۶ | ۲,۱۷ |
| ۴۷ | ۱۳۹۲/۴/۷ | ۱,۶۰ | ۷۴,۹۴ | ۹۹۳,۳۹ | ۳۲,۲۴ | ۲,۸۸ |
| ۴۸ | ۱۳۹۲/۴/۸ | ۱,۲۰ | ۷۸,۴۹ | ۹۹۳,۸۷ | ۳۲,۳۴ | ۲,۹۲ |
| ۴۹ | ۱۳۹۲/۴/۹ | ۰,۹۰ | ۶۹,۴۷ | ۹۹۴,۷۱ | ۳۲,۶۴ | ۲,۱۳ |
| ۵۰ | ۱۳۹۲/۴/۱۰ | ۰,۴۴ | ۵۰,۰۸ | ۹۹۵,۰۸ | ۳۲,۶۳ | ۱,۶۱ |
| ۵۱ | ۱۳۹۲/۴/۲۴ | ۰,۷۴ | ۷۲,۶۰ | ۹۹۵,۰۸ | ۳۴,۷۸ | ۲,۶۲ |
| ۵۲ | ۱۳۹۲/۴/۲۵ | ۱,۰۷ | ۷۲,۵۳ | ۹۹۴,۲۱ | ۳۴,۷۱ | ۳,۰۴ |
| ۵۳ | ۱۳۹۲/۴/۲۶ | ۰,۸۵ | ۷۲,۴۸ | ۹۹۳,۸۸ | ۳۴,۵۶ | ۲,۳۰ |
| ۵۴ | ۱۳۹۲/۴/۲۷ | ۰,۸۳ | ۷۲,۸۰ | ۹۹۳,۳۸ | ۳۴,۷۱ | ۱,۹۳ |
| ۵۵ | ۱۳۹۲/۴/۲۸ | ۰,۶۵ | ۷۱,۲۹ | ۹۹۱,۵۴ | ۳۴,۹۲ | ۲,۳۹ |
| ۵۶ | ۱۳۹۲/۴/۲۹ | ۰,۷۹ | ۷۱,۰۰ | ۹۹۱,۱۳ | ۳۴,۹۶ | ۲,۷۱ |
| ۵۷ | ۱۳۹۲/۴/۳۰ | ۱,۵۸ | ۷۳,۸۶ | ۹۹۲,۰۴ | ۳۴,۹۶ | ۲,۴۵ |
| ۵۸ | ۱۳۹۲/۴/۳۱ | ۱,۱۶ | ۷۰,۶۳ | ۹۹۱,۸۱ | ۳۴,۹۶ | ۲,۰۹ |
| ۵۹ | ۱۳۹۲/۵/۱ | ۱,۵۹ | ۶۸,۲۶ | ۹۹۲,۹۸ | ۳۴,۹۶ | ۲,۳۴ |

- پایش فعالیت: به عنوان مثال شناسایی کلاهبرداری بوسیله گوشی‌های همراه با پایش نحوه فعالیت گوشی و یا خرید و فروش‌های مشکوک تجاری [۶].
- تشخیص عیب: به عنوان مثال پایش فعالیت‌ها برای شناسایی بروز مشکل در موتورها، ژنراتورها، خطوط لوله، پایش سازه‌های صنعتی مثلاً شناسایی وجود ترک در دیوارها یا تجهیزات فضایی در شاتل‌ها [۷، ۸، ۹].
- تحلیل تصاویر ماهواره‌ای: شناسایی پدیده‌های جدید و یا پدیده‌هایی که تا کنون دسته‌بندی نشده‌اند [۱۰، ۱۱].
- شناسایی موارد جدید در تصاویر: به عنوان مثال استفاده از آنها در سیستم‌های تجسسی [۱۲].
- شناسایی پدیده‌های متحرک: شناسایی پدیده‌هایی که نسبت به زمینه تصویر متحرک هستند، به عنوان مثال شناسایی ورود افراد به بانک [۱۳].
- پایش سری‌های زمانی: به عنوان مثال پایش ایمنی در امور مهمی مثل حفاری و یافتن الگوهای غیر عادی در سری‌های زمانی [۱۴].
- پایش وضعیت در امور پزشکی: مثلاً پایش میزان ضربان قلب بیمار [۱۵، ۱۶].
- شناسایی متن: به عنوان مثال شناسایی عناوین در اسناد خبری [۱۷].

روش ضریب داده پرت محلی ، یکی ازمهمترین روشهای داده کاوی است. این روش بسیار قدرتمند در شناسایی ناهنجاری‌های داده‌ها در یادگیری ماشین عمل می کند [۱۸] . در شماری از کاربردهای عملی مثل تشخیص تقلب در کارت‌های اعتباری [۱۹]، بازاریابی محصولات [۲۰]، امنیت شبکه [۲۱] و شناسایی خطاهای موجود در برداشت داده‌های آزمایشگاهی [۲۲] نیزکاربرد ویژه دارد.

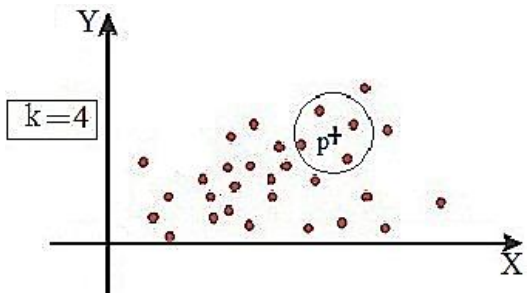
روش ضریب داده پرت محلی از مفهوم داده پرت محلی استفاده می‌کند. به این ترتیب که به هر داده درجه‌ای با توجه به چگالی همسایه‌های محلی آن نسبت می‌دهد. این درجه ضریب پرت بودن داده (LOF) نامیده می‌شود. ضریب LOF به میزان ایزوله بودن یک داده وقتی که با همسایه‌های محلی آن مقایسه می‌شود، بستگی دارد. در الگوریتم LOF تفاوت در چگالی بین یک داده و همسایه‌های آن، درجه پرت بودن را مشخص می‌کند. برای هر داده می‌توان مقدار ضریب LOF را محاسبه کرد که میزان پرت بودن آن را نشان می‌دهد. به طور شهودی مقادیر بالای ضریب LOF می‌تواند نماینده داده‌های مشکوک به خطا باشد؛ در حالی که مقدار پایین آن نشان دهنده نرمال بودن یک داده است، چون دلالت بر چگالی پایین همسایگی آن داده دارد.

برای محاسبه مقادیر LOF تمامی اعضای یک مجموعه داده، احتیاج به یافتن k تا همسایه‌های نزدیک هر داده است.

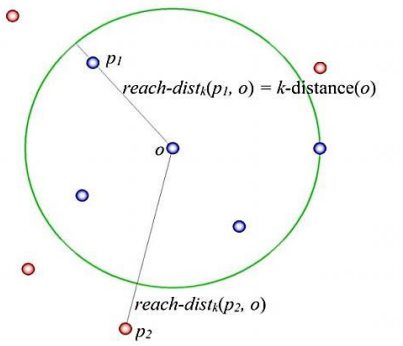
برای یافتن این همسایه‌ها از الگوریتم k تا نزدیکترین همسایه‌ها[1] (kNN) استفاده می‌شود. قبل از تشریح الگوریتم، پارامترهای مورد استفاده معرفی می‌شوند. D مجموعه داده، o، p و q چند داده نمونه که در D واقع شده‌اند و k یک عدد صحیح مثبت است که توسط کاربر انتخاب می‌شود. d(p,q) تابع اندازه‌گیری فاصله بین داده‌های p و q است که می‌تواند هر تابع اندازه‌گیری فاصله باشد. برای محاسبه kNN داده p، ابتدا فاصله آن با تمامی داده‌های موجود در مجموعه داده D بدست آید. پس از مرتب کردن فاصله‌های محاسبه شده به صورت صعودی، k تا داده‌ای که در فاصله کمتری نسبت به p واقع شده‌اند، در k تا فاصله همسایگی آن قرار می‌گیرند، که k تا همسایه محلی آن نامیده می‌شوند. به عنوان مثال، شکل (۲) نشان دهنده ۴ همسایگی داده p، از مجموعه داده D است.

جهت محاسبه LOF داده p که در مجموعه داده D قرار دارد، انجام گام‌های زیر ضروری است:

گام اول: محاسبه k امین فاصله p [2]




**شکل (۲) kNN داده p، به ازای k=۴.**




**شکل (۳) مثالی از چگالی دسترسی به ازای k=۴.**

برای هر داده p، k-distance(p)، فاصله k امین همسایه نزدیک p از آن است. برای محاسبه k-distance(p) می‌توان از الگوریتم kNN استفاده کرد.

به این ترتیب که ابتدا فاصله k تا همسایه نزدیک p را پیدا کرد، سپس k امین فاصله را به عنوان k-distance(p) انتخاب کرد. داده‌هایی که در k تا فاصله همسایگی p واقع می‌شوند دارای دو ویژگی هستند:

الف) برای حداقل k تا داده $d(p, o') \le d(p, o), o' \in D \setminus \{p\}$

ب) برای حداکثر k-1 داده $o' \in D \setminus \{p\}, d(p, o') < d(p, o)$

بنابراین k-distance(p) تخمینی از چگالی همسایگی اطراف p را مشخص می‌کند.

گام دوم: پیدا کردن همسایگی k امین فاصله p [3]

هر داده‌ای که فاصله آن از p، از k-distance(p) کمتر و یا مساوی باشد، در k امین فاصله همسایگی p قرار می‌گیرد. این تعریف به صورت زیر ارائه می‌شود:

$$N_{k\text{-distance}(p)}(p) = \{q \in D \setminus \{p\} \mid d(p, q) \le k\text{-distance}(p)\}$$

گام سوم: محاسبه چگالی دسترسی داده p نسبت به داده o

برای هر داده o در k تا فاصله همسایگی p، چگالی دسترسی[4] داده p نسبت به داده o به صورت زیر تعریف می‌شود:

$$reachdist_k(p, o) = \max\{k\text{-distance}(o), d(p, o)\}$$

شکل (۳) نشان دهنده مثالی از فاصله دسترسی به ازای k=۴ است. همان‌طور که در این شکل نشان داده شده است، اگر داده p خارج از k-distance(o) واقع شده باشد ($p_2$ در شکل)، چگالی دسترسی فاصله بین آنها می‌شود یعنی $d(o,p_2)$. حال اگر فاصله آنها از k-distance(o) کمتر باشد ($p_1$ در شکل)، چگالی دسترسی برابر k-distance(o) است.

گام چهارم: محاسبه چگالی دسترسی محلی p[5]

چگالی دسترسی محلی p، معکوس میانگین چگالی دسترسی k تا همسایه‌های نزدیک p است:

$$lrd_k(p) = \left[\frac{|N_k(p)|}{\sum_{o \in N_{k(p)}} reach\text{-}dist_k(p,o)}\right] \qquad (۱)$$

برای محاسبه چگالی داده p از چگالی دسترسی محلی آن استفاده می‌شود.

با استفاده از چگالی دسترسی محلی می‌توان ضریب LOF را محاسبه کرد.

گام پنجم: محاسبه ضریب داده پرت محلی (LOF)

[1] k–Nearest Neighbor (kNN)
[2] k–Distance of p
[3] k–Distance Neighborhood of p
[4] Reachability Distance of p w.r.t Object o
[5] Local Reachability Density of p

از ضریب LOF برای تشخیص نرمال و یا عدم نرمال بودن یک داده استفاده می‌شود. LOF(p) میانگین نسبت‌های چگالی دسترسی محلی p و k تا همسایه‌های آن است:

$$LOF_k(p) = \frac{\sum_{o\epsilon N_k(p)} \frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \quad (۲)$$

در ادامه برخی از توابع اندازه‌گیری فاصله که در این تحقیق در الگوریتم‌های kNN و LOF بکار رفته‌اند، به طور مختصر معرفی شده است.

### ۳-۲- توابع اندازه‌گیری فاصله

ماتریس $X_{mx\times n}$ داده شده است، می‌توان آن را به عنوان $mx$ تا بردار سطری $(1\times n)$ با عناصر $x_1, x_2, x_3, \dots, x_{m_x}$ در نظر گرفت. همچنین ماتریس $Y_{my\times n}$ که می‌توان آن را به عنوان $my$ تا بردار سطری $(1\times n)$ با عناصر $y_1, y_2, y_3, \dots, y_{m_y}$ در نظر گرفت. توابع مختلف اندازه‌گیری فاصله بین عناصر بردار $x_s$ و $y_t$ به صورت زیر تعریف می‌شوند:

۱- روش مینکوسکی[1]

$$d_{st} = \sqrt[p]{\sum_{j=1}^{n} |x_{sj} - y_{tj}|^p} \quad (۳)$$

p مرتبه فاصله مینکوسکی نامیده می‌شود که می‌تواند از ۱ تا بینهایت متغیر باشد. فاصله مینکوسکی برای حالت خاص p=۱ فاصله بلاک شهری، برای p=۲ فاصله اقلیدسی و برای حالت خاص p= ∞، فاصله چبیشف را نتیجه می‌دهد. $n$ تعداد متغیرها، همچنین $x_{sj}$ و $y_{tj}$ به ترتیب مقادیر $j$ امین متغیر در نقاط $x_s$ و $y_t$ است.

۲- روش اقلیدسی[2]

فاصله اقلیدسی حالت خاصی از فاصله مینکوسکی با p=۲ است:

$$d_{st}^2 = (x_s - y_t)(x_s - y_t)' \quad (۴)$$

۳- روش بلاک شهری[3]

فاصله بلاک شهری حالت خاص p=∞ در رابطه مینکوسکی است:

$$d_{st} = \sum_{j=1}^{n} |x_{sj} - y_{tj}| \quad (۵)$$

۴- روش ماهالانوبیس[4]

$$d_{st}^2 = (x_s - y_t)C^{-1}(x_s - y_t)' \quad (۶)$$

در این رابطه $C$ کوواریانس ماتریس ورودی است.

۵- روش چبیشف[5]

$$d_{st} = max_j\{|x_{sj} - y_{tj}|\} \quad (۷)$$

۶- روش اقلیدسی استاندارد شده[6]

$$d_{st}^2 = (x_s - y_t)V^{-1}(x_s - y_t)' \quad (۸)$$

V یک ماتریس قطری n در n است که j امین المان قطری آن $S(j)^2$ است، که S برداری است که شامل وزن‌های معکوس[7] است.

۷- فاصله کسینوس[8]

$$d_{st} = \left(1 - \frac{x_s - y_t'}{\sqrt{(x_s x_s')(y_t y_t')}}\right) \quad (۹)$$

۸- فاصله همبستگی[9]

$$d_{st} = \left(1 - \frac{(x_s - \bar{x}_s)(y_t - \bar{y}_t)'}{\sqrt{(x_s - \bar{x}_s)(x_s - \bar{x}_s)'}\sqrt{(y_t - \bar{y}_t)(y_t - \bar{y}_t)'}}\right) \quad (۱۰)$$

که

$$\bar{x}_s = \frac{1}{n}\sum_j x_{sj} \quad (۱۱)$$

$$\bar{y}_t = \frac{1}{n}\sum_j y_{tj} \quad (۱۲)$$

۹- فاصله هَمینگ[10]

$$d_{st} = \left(\frac{\#(x_{sj} \neq y_{tj})}{n}\right) \quad (۱۳)$$

۱۰- فاصله جاکارد[11]

$$d_{st} = \frac{\#\left[(x_{sj} \neq y_{tj}) \cap \left((x_{sj} \neq 0) \cup (y_{tj} \neq 0)\right)\right]}{\#[(x_{sj} \neq 0) \cup (y_{tj} \neq 0)]} \quad (۱۴)$$

۱۱- فاصله اسپیرمَن[12]

$$d_{st} = 1 - \frac{(r_s - \bar{r}_s)(r_t - \bar{r}_t)'}{\sqrt{(r_s - \bar{r}_s)(r_s - \bar{r}_s)'}\sqrt{(r_t - \bar{r}_t)(r_t - \bar{r}_t)'}} \quad (۱۵)$$

که

$r_{sj}$ رنک[13] $x_{sj}$ می‌باشد که روی $x_{1j}, x_{2j}, \dots, x_{mx,j}$ گرفته شده است. $r_{tj}$ رنک $y_{tj}$ است که روی $y_{1j}, y_{2j}, \dots, y_{mx,j}$ گرفته شده است. $r_s$ و $r_t$ رنک بردارهای $x_s$ و $y_t$ است، یعنی $r_s = (r_{s_1}, r_{s_2}, \dots, r_{s_n})$ و $r_t = (r_{t_1}, r_{t_2}, \dots, r_{t_n})$.

$$\bar{r}_s = \frac{1}{n}\sum_j r_{sj} = \frac{n+1}{2} \quad (۱۶)$$

[1] Minkowski Metric
[2] Euclidean Distance
[3] City Block Metric
[4] Mahalanobis Distance
[5] Chebychev Distance
[6] Standardized Euclidean Distance
[7] Inverse Weights
[8] Cosine Distance
[9] Correlation Distance
[10] Hamming Distance
[11] Jaccard Distance
[12] Spearman Distance
[13] Rank

$$\bar{r}_t = \frac{1}{n}\sum_j r_{tj} = \frac{n+1}{2} \quad (۱۷)$$

## ۴- نتایج

در این بخش، نتایج آزمون LOF به ازای توابع مختلف اندازه‌گیری فاصله روی نمونه مذکور ارائه شده است. جهت محاسبه ضریب LOF برای هر داده از پارامتر k استفاده می‌شود. مقدار k وابسته به مسأله مورد نظر و طبیعت داده‌هایی است که قرار است داده‌های مشکوک به خطا از بین آنها شناسایی شود. به طور کلی تمامی الگوریتم‌های شناسایی داده‌های پرت وابسته به یک یا چند پارامتر هستند که توسط کاربر و بر حسب نوع مسئله و شرایط فیزیکی داده‌ها انتخاب می‌شوند. تا کنون هیچ قانون مشخص و پذیرفته شده‌ای برای انتخاب این پارامترها ارائه نشده است و اگر ارائه شده باشد نیز نسبی است و در تمامی حالات معتبر نیست.

در الگوریتم نزدیک‌ترین همسایه نیز این حالت وجود دارد و انتخاب مقدار k از روی سعی و خطا و بسته به شرایط مسئله است. در این الگوریتم به طور کلی یک قانون برقرار است. اگر تعداد همسایه‌های انتخابی به تعداد قابل توجهی کم انتخاب شود آنگاه دقت روش بسیار پایین خواهد آمد در غیر این صورت اگر تعداد همسایه‌های انتخابی از یک حد مشخص بالاتر انتخاب شود آنگاه دقت روش تغییری نخواهد کرد و فقط حجم محاسبات و زمان اجرای برنامه کامپیوتری افزایش خواهد یافت. در این تحقیق، با توجه به اینکه هر داده برداشت شده مربوط به یک روز است، داده‌های برداشت شده طی یک هفته به عنوان داده‌های همسایه در نظر گرفته شده‌اند. بنابراین مقدار k برابر ۷ انتخاب شده است؛ این بدان معنی است که برای هر داده ۷ همسایه نزدیک آن به عنوان حوزه همسایگی آن داده تعیین شده است.

پس از محاسبه ضریب LOF برای هر داده، باید بر اساس ضریب محاسبه شده در مورد نرمال و یا نرمال نبودن آن داده تصمیم‌گیری شود. با توجه به فرمول LOF اگر همه داده‌ها به صورت مرتب و با فاصله دقیقاً یکسانی در صفحه قرار گیرند، آنگاه ضریب LOF همه داده‌ها (به جز داده‌های مرزی) ۱ خواهد بود. از طرفی اگر در اطراف یک داده فشردگی ایجاد شود، یعنی چگالی حول آن داده بیش از سایر داده‌ها شود، آنگاه ضریب LOF مربوط به آن داده از یک کمتر می‌شود و هرچه این چگالی بیشتر باشد این ضریب به صفر نزدیکتر می‌شود. همچنین اگر حول داده‌ای چگالی کمتر از میزان چگالی یکنواخت بقیه داده‌ها باشد، LOF آن داده از ۱ بیشتر می‌شود و هر چه تراکم داده‌ها حول آن داده کمتر باشد LOF عدد بزرگتری می‌شود و حتی می‌تواند بینهایت شود. بنابراین بسته به مسأله مورد نظر باید عددی بزرگتر از ۱ به عنوان آستانه تشخیص انتخاب شود. در ا ین آزمایشات عدد ۱,۳ به عنوان آستانه انتخاب شده است. این بدان معنی است که داده‌هایی که چگالی اطراف آن‌ها بیشتر از ۳۰ درصد از چگالی یکنواخت داده‌ها کمتر است به عنوان داده‌های مشکوک به خطا در نظر گرفته می‌شوند. با توجه به رابطه LOF می‌توان دید که اگر مقدار آستانه از ۱,۳ بیشتر انتخاب شود برخی از داده‌ها از حوزه داده‌های خطا خارج شده و داده‌های نرمال به حساب خواهند آمد؛ همچنین اگر این آستانه از ۱,۳ کمتر انتخاب شود برخی از داده‌های نرمال وارد حوزه داده‌های خطا شده و به عنوان داده خطا شناخته خواهند شد.

بنابراین در این مطالعه با توجه به توزیع داده‌ها در صفحه دو بعدی و فاصله نسبی نقاط از هم و با توجه به دهک‌های بالا و پایین فواصل داده‌ها از میانگین، مقدار پارامتر آستانه ۱,۳ انتخاب شده است. نتایج آزمون LOF روی نمونه‌ها به ازای k=۷ و d=۱,۳، و بر حسب توابع مختلف اندازه‌گیری فاصله در جداول ۲ تا ۶ ارائه شده است. برای انجام تمامی آزمون‌ها یک کد کامپیوتری در محیط نرم‌افزار MATLAB نوشته شد. همچنین جهت یکسان‌سازی شرایط آزمون از لحاظ زمان اجرا، تمامی آزمون‌ها بر روی یک سیستم با واحد پردازنده مرکزی ۲GHz، رم ۳GB و سیستم عامل ویندوز ۷ (۶۴ بیتی) انجام شده است. زمان اجرا، مدت زمان مورد نیاز جهت اجرای روش توسط کامپیوتر را مشخص می‌کند. هرچه زمان اجرای روش LOF با یک تابع اندازه‌گیری فاصله مشخص کمتر باشد، می‌توان در مدت زمان کمتری به نتایج دست یافت. از اینرو از هزینه‌های زمانی کاسته می‌شود.

همانطور که از نتایج موجود در جداول ۲ تا ۶ مشخص است، روش LOF به ازای توابع اقلیدسی، اقلیدسی استاندارد شده، بلاک شهری، مینکوسکی، چبیشف و ماهالانوبیس روی نمونه‌های مختلف، نتایج کاملاً مشابهی ارائه کرده است. و تنها اختلاف بین آنها، در زمان اجرای روش حین بکارگیری با توابع مختلف اندازه‌گیری فاصله است. بنابراین جهت شناسایی داده‌های مشکوک به خطا در داده‌های گردآوری

شده از بویه‌ها، استفاده از هر یک از این توابع سبب تغییر نتایج نمی‌شود و می‌توان به صورت دلخواه از هر یک از توابع مذکور استفاده کرد. در اینجا اختلاف زمان اجرای توابع مختلف بسیار ناچیز بوده و قابل چشم‌پوشی است؛ اما هنگام کار روی نمونه‌های بسیار بزرگ، زمان اجرا حائز اهمیت است و می‌تواند سبب صرفه‌جویی در هزینه‌های زمانی و مالی شود. همچنین نتایج اجرای روش LOF بر حسب تابع مینکوسکی و با تغییر مقادیر p تغییر نکرده است.

اما همواره نمی‌توان گفت تغییر مقادیر p سبب تغییر نتایج نشده و می‌توان از یک مقدار p دلخواه استفاده کرد. مثلاً در نمونه سرعت باد اگر مقدار p بیشتر از ۳۶۵ انتخاب شود، مقادیر کلی LOF داده‌ها بیشتر شده و در نتیجه تعداد داده‌های بیشتری به عنوان کاندیدای مشکوک به خطا انتخاب خواهند شد. به عنوان مثال در شکل (۴)، مقادیر ضریب LOF داده‌های میزان سرعت باد به ازای p=۴۰ و p=۵۰۰ نشان داده شده است. همانطور که از این شکل مشخص است، مقادیر LOF داده‌ها به ازای p=۵۰۰ نسبت به p=۴۰ به طور قابل توجهی بیشتر است. اما در همه نمونه‌ها روش LOF بر حسب توابع کسینوس، همبستگی و اسپیرمَن هیچ نتیجه‌ای ارائه نکرده و نتوانسته است با موفقیت اجرا شود.

بنابراین بکارگیری این توابع با روش LOF جهت شناسایی داده‌های مشکوک به خطا در نمونه‌های مذکور پیشنهاد نمی‌شود. همچنین بکارگیری روش LOF با توابع هَمینگ و جاکارد سبب شده تا تمامی داده‌ها به عنوان داده مشکوک به خطا تشخیص داده شوند. این امر به این دلیل است که ضریب LOF تمامی داده‌ها با استفاده از این توابع برابر ۷,۳۷۵۰ شده است. چون این مقدار از مقدار پارامتر آستانه بیشتر است، در نتیجه تمامی داده‌ها به عنوان داده مشکوک به خطا تشخیص داده شده‌اند. بنابراین می‌توان گفت بکارگیری این دو تابع جهت شناسایی داده‌های مشکوک به خطا در نمونه‌های مذکور نیز سبب نتایج اشتباه شده و استفاده از آنها با روش LOF پیشنهاد نمی‌شود.

با توجه به نتایج جدول (۲) داده‌های با اندیس ۱، ۱۱ و ۱۲ به عنوان کاندیدای مشکوک به خطا تشخیص داده شده‌اند. در شکل (۵)، این داده‌ها با یک علامت دایره به دور آنها مشخص شده‌اند. با توجه به فیزیک مسئله این داده‌ها خطا نیستند، چون این اندیس‌ها به ترتیب مربوط به داده‌های ۰,۳۸، ۰,۳۷ و ۰,۳۷ متر است و این مقادیر به عنوان ارتفاع موج شاخص طبیعی است. و تنها علت انتخاب این داده‌ها، دور افتادگی نسبی آنها از اعضای نمونه است. طبق نتایج جدول (۳)، هیچ داده‌ای به عنوان مشکوک به خطا شناسایی نشده است. با توجه به اینکه بیشتر توابع بر عدم وجود داده خطا در این نمونه تأکید دارند، بنابراین می‌توان گفت این نمونه فاقد خطا بوده و نتایج حاصل شده از بویه قابل اتکا است.

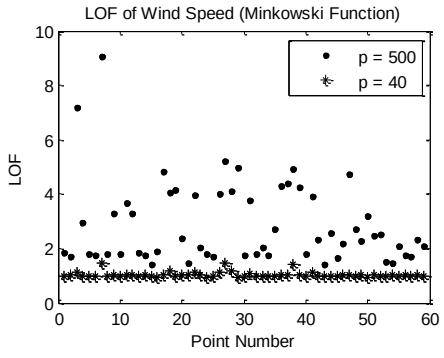
در جدول (۴)، داده‌های با اندیس ۲ و ۳ به عنوان کاندیدای مشکوک به خطا تشخیص داده شده‌اند. در شکل (۶)، این داده‌ها با یک علامت دایره به دور آنها مشخص شده‌اند. با توجه به فیزیک مسئله این داده‌ها نیز خطا نیستند، چون این اندیس‌ها به ترتیب مربوط به داده‌های با مقادیر ۱۰۰۵,۴۷ و ۱۰۰۵,۴۹ اتمسفر بوده و این مقادیر به عنوان فشارهوا طبیعی است. در شکل (۷) و (۸)، داده‌های مشکوک به خطای شناسایی شده در نمونه‌های مربوط به دمای آب و سرعت باد نشان داده شده است. در این نمونه‌ها نیز داده‌های علامت‌گذاری شده تفاوت چندانی با سایر اعضای نمونه نداشته و بر اساس فیزیک مسئله، این داده‌ها نمی‌توانند کاندیدای داده خطا باشند و صرفاً بر اساس تفاوت ناچیزی که با الگوی نرمال جامعه آماری دارند، توسط روش به عنوان کاندیدای خطا انتخاب شده‌اند. بنابراین می‌توان گفت در این مورد نیز داده‌های برداشت شده از بویه قابل اتکا بوده و می‌توان از آنها به منظور اهداف مختلف استفاده کرد. با توجه به نتایج موجود در جداول، در نمونه‌های مختلف داده‌هایی به عنوان کاندیدای مشکوک به خطا انتخاب شدند. و مشاهده شد که با توجه به فیزیک مسئله این داده‌ها نمی‌توانند خطا باشند. بنابراین می‌توان گفت داده‌هایی که به عنوان خطا توسط روش انتخاب می‌شوند، همواره به این معنا نیست که واقعاً خطایی حین گردآوری داده‌ها رخ داده و چه بسا ممکن است بر حسب شرایط خاص اندازه‌گیری به وجود آمده باشند و گویای اطلاعات بسیار مهمی در خصوص سیستم اندازه‌گیری بوده که تا کنون به آنها پرداخته نشده است و یا در حالت‌های بسیار نادر رخ خواهند داد. مثلاً وجود یک داده متناقض در نمونه میزان سرعت باد می‌تواند بیانگر وقوع طوفان‌های شدید در برخی از اوقات سال باشد.

**جدول (۲) داده‌های مشکوک به خطا شناسایی شده توسط روش LOF، روی نمونه ارتفاع موج شاخص (k=۷ و d=۱,۳).**

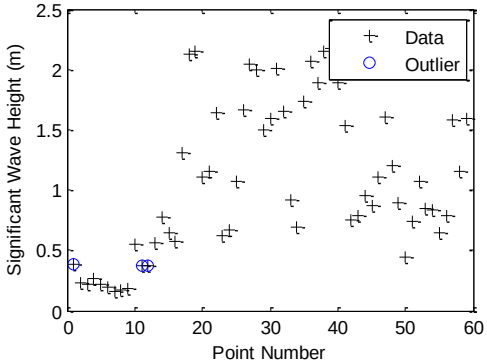
| تابع اندازه‌گیری فاصله | شماره (اندیس) داده‌های مشکوک به خطا شناسایی شده | تعداد | زمان اجرا (ثانیه) |
|---|---|---|---|
| اقلیدسی | ۱، ۱۱، ۱۲ | ۳ | ۰,۴۲۲۴۷۲ |
| اقلیدسی استاندارد شده | ۱، ۱۱، ۱۲ | ۳ | ۰,۵۴۴۲۲۲ |
| بلاک شهری | ۱، ۱۱، ۱۲ | ۳ | ۰,۳۲۳۶۴۲ |
| مینکوسکی (p= ۵) | ۱، ۱۱، ۱۲ | ۳ | ۰,۳۳۲۰۵۰ |
| مینکوسکی (p= ۱۰) | ۱، ۱۱، ۱۲ | ۳ | ۰,۳۳۲۰۰۶ |
| مینکوسکی (p= ۲۰) | ۱، ۱۱، ۱۲ | ۳ | ۰,۳۳۲۵۵۱ |
| مینکوسکی (p= ۴۰) | ۱، ۱۱، ۱۲ | ۳ | ۰,۳۳۶۰۰۱ |
| چبیشف | ۱، ۱۱، ۱۲ | ۳ | ۰,۳۲۳۹۷۴ |
| ماهالانوبیس | ۱، ۱۱، ۱۲ | ۳ | ۰,۶۷۷۹۲۵ |
| کسینوس | بدون عملکرد | ۰ | ۰,۳۹۴۳۵۸ |
| همبستگی | بدون عملکرد | ۰ | بدون عملکرد |
| اسپیرمَن | بدون عملکرد | ۰ | ۰,۴۶۵۰۴۴ |
| هَمینگ | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۸۶۵۸۴ |
| جاکارد | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۸۸۲۲۱ |

**جدول (۳) داده‌های مشکوک به خطا شناسایی شده توسط روش LOF، روی نمونه میزان رطوبت (k=۷ و d=۱,۳).**

| تابع اندازه‌گیری فاصله | شماره (اندیس) داده‌های مشکوک به خطا شناسایی شده | تعداد | زمان اجرا (ثانیه) |
|---|---|---|---|
| اقلیدسی | --- | ۰ | ۰,۳۳۵۰۹۸ |
| اقلیدسی استاندارد شده | --- | ۰ | ۰,۳۵۳۸۸۳ |
| بلاک شهری | --- | ۰ | ۰,۳۱۳۲۳۱ |
| مینکوسکی (p= ۵) | --- | ۰ | ۰,۳۲۴۹۳۱ |
| مینکوسکی (p= ۱۰) | --- | ۰ | ۰,۳۳۳۶۰۸ |
| مینکوسکی (p= ۲۰) | --- | ۰ | ۰,۳۲۷۸۹۴ |
| مینکوسکی (p= ۴۰) | --- | ۰ | ۰,۳۲۵۳۶۴ |
| چبیشف | --- | ۰ | ۰,۳۱۲۴۵۰ |
| ماهالانوبیس | --- | ۰ | ۰,۳۶۲۵۸۰ |
| کسینوس | بدون عملکرد | ۰ | ۰,۳۹۱۹۵۵ |
| همبستگی | بدون عملکرد | ۰ | بدون عملکرد |
| اسپیرمَن | بدون عملکرد | ۰ | ۰,۴۲۸۸۹۷ |
| هَمینگ | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۹۳۴۶۴ |
| جاکارد | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۸۷۰۶۹ |




**شکل (۴) مقادیر ضریب LOF نمونه سرعت باد با تابع مینکوسکی.**



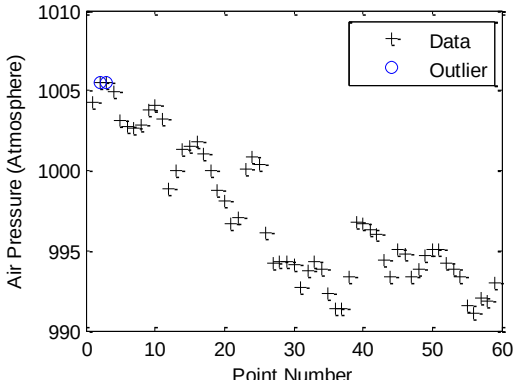

**شکل (۵) نمونه ارتفاع موج شاخص با داده‌های مشکوک به خطا.**

**جدول (۴) داده‌های مشکوک به خطا شناسایی شده توسط روش LOF، روی نمونه میزان فشار هوا (k= ۷ و d= ۱,۳).**

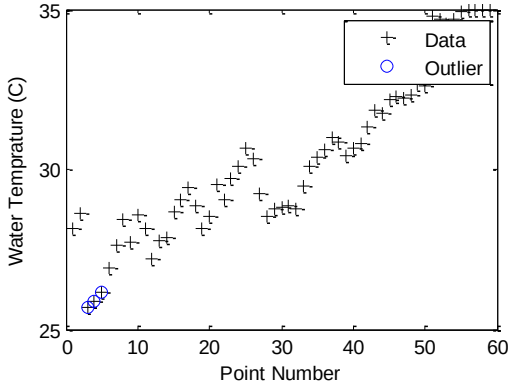
| تابع اندازه‌گیری فاصله | شماره (اندیس) داده‌های مشکوک به خطا شناسایی شده | تعداد | زمان اجرا (ثانیه) |
|---|---|---|---|
| اقلیدسی | ۳ ،۲ | ۲ | ۰,۲۰۰۰۶۲ |
| اقلیدسی استاندارد شده | ۳ ،۲ | ۲ | ۰,۳۷۱۹۷۴ |
| بلاک شهری | ۳ ،۲ | ۲ | ۰,۳۳۵۰۷۷ |
| مینکوسکی (p= ۵) | ۳ ،۲ | ۲ | ۰,۳۴۱۰۹۵ |
| مینکوسکی (p= ۱۰) | ۳ ،۲ | ۲ | ۰,۳۳۹۳۷۷ |
| مینکوسکی (p= ۲۰) | ۳ ،۲ | ۲ | ۰,۳۴۴۳۸۴ |
| مینکوسکی (p= ۴۰) | ۳ ،۲ | ۲ | ۰,۳۴۳۰۴۴ |
| چبیشف | ۳ ،۲ | ۲ | ۰,۳۲۷۶۳۷ |
| ماهالانوبیس | ۳ ،۲ | ۲ | ۰,۳۸۳۳۹۳ |
| کسینوس | بدون عملکرد | ۰ | ۰,۴۲۰۶۷۷ |
| همبستگی | بدون عملکرد | ۰ | بدون عملکرد |
| اسپیرمَن | بدون عملکرد | ۰ | ۰,۴۳۲۸۶۵ |
| هَمینگ | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۹۹۲۸۹ |
| جاکارد | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۹۵۳۳۵ |

**جدول (۵) داده‌های مشکوک به خطا شناسایی شده توسط روش LOF، روی نمونه دمای آب (k= ۷ و d= ۱,۳).**

| تابع اندازه‌گیری فاصله | شماره (اندیس) داده‌های مشکوک به خطا شناسایی شده | تعداد | زمان اجرا (ثانیه) |
|---|---|---|---|
| اقلیدسی | ۳، ۴، ۵ | ۳ | ۰,۳۴۱۲۱۲ |
| اقلیدسی استاندارد شده | ۳، ۴، ۵ | ۳ | ۰,۳۷۰۲۴۳ |
| بلاک شهری | ۳، ۴، ۵ | ۳ | ۰,۳۳۱۰۶۰ |
| مینکوسکی (p= ۵) | ۳، ۴، ۵ | ۳ | ۰,۳۴۵۷۱۸ |
| مینکوسکی (p= ۱۰) | ۳، ۴، ۵ | ۳ | ۰,۳۴۳۴۴۶ |
| مینکوسکی (p= ۲۰) | ۳، ۴، ۵ | ۳ | ۰,۳۴۲۲۵۷ |
| مینکوسکی (p= ۴۰) | ۳، ۴، ۵ | ۳ | ۰,۳۴۱۸۵۸ |
| چبیشف | ۳، ۴، ۵ | ۳ | ۰,۳۳۱۳۷۴ |
| ماهالانوبیس | ۳، ۴، ۵ | ۳ | ۰,۳۸۲۴۸۷ |
| کسینوس | بدون عملکرد | ۰ | ۰,۳۹۸۴۳۵ |
| همبستگی | بدون عملکرد | ۰ | بدون عملکرد |
| اسپیرمَن | بدون عملکرد | ۰ | ۰,۴۳۹۳۳۵ |
| هَمینگ | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۹۵۰۳۲ |
| جاکارد | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۹۶۷۲۶ |



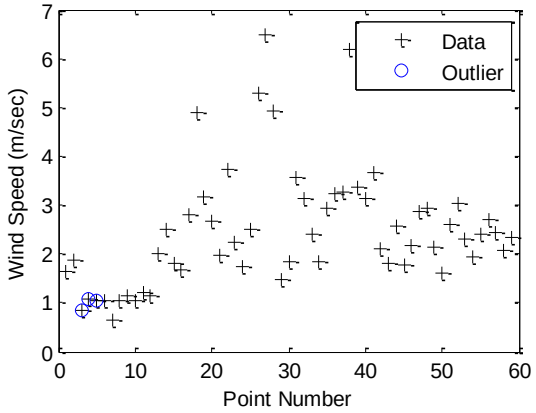

**شکل (۶) نمونه میزان فشار هوا با داده‌های مشکوک به خطا.**




**شکل (۷) نمونه میزان دمای آب با داده‌های مشکوک به خطا.**

**جدول (۶) داده‌های مشکوک به خطا شناسایی شده توسط روش LOF، روی نمونه میزان سرعت باد (۷ =k و ۱,۳ =d).**

| تابع اندازه‌گیری فاصله | شماره (اندیس) داده‌های مشکوک به خطا شناسایی شده | تعداد | زمان اجرا (ثانیه) |
|---|---|---|---|
| اقلیدسی | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۳۱۲۲۷ |
| اقلیدسی استاندارد شده | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۷۱۵۶۷ |
| بلاک شهری | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۲۵۳۶۰ |
| مینکوسکی (p= ۵) | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۳۸۲۳ |
| مینکوسکی (p= ۱۰) | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۴۰۵۹۵ |
| مینکوسکی (p= ۲۰) | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۴۱۸۲۳ |
| مینکوسکی (p= ۴۰) | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۵۴۶۳۰ |
| چبیشف | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۲۷۹۶۳ |
| ماهالانوبیس | ۷، ۲۷، ۳۸ | ۳ | ۰,۳۷۶۵۸۸ |
| کسینوس | بدون عملکرد | ۰ | ۰,۴۰۱۲۴۱ |
| همبستگی | بدون عملکرد | ۰ | بدون عملکرد |
| اسپیرمَن | بدون عملکرد | ۰ | ۰,۴۳۲۵۰۱ |
| هَمینگ | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۹۸۴۰۲ |
| جاکارد | ۱، ۲، ۳، ۴، ۵، ...، ۵۵، ۵۶، ۵۷، ۵۸، ۵۹ | ۵۹ | ۰,۳۹۶۰۰۱ |




**شکل (۸) نمونه میزان سرعت باد با داده‌های مشکوک به خطا.**

از اینرو توجه مسئولین و مهندسان می‌تواند به آن جلب شود تا در صورت وقوع دوباره آن، تدابیر لازم را بکار گیرند. بنابراین می‌توان گفت که از تمامی این روش‌ها فقط به منظور تسهیل در شناسایی داده‌های مشکوک به خطا استفاده می‌شود و تشخیص نهایی خطا و یا صحیح بودن داده بر عهده فرد متخصص است. اوست که با توجه به ماهیت مسئله و شرایط داده‌ها، تصمیم‌گیری نهایی در خصوص خطا و یا نرمال بودن آن را انجام می‌دهد.

## ۵- نتیجه‌گیری

حین انجام آزمایش‌های مختلف، عوامل متعددی سبب بروز خطا در اندازه‌گیری‌ها می‌شوند. شناسایی این خطاها و در صورت امکان برطرف کردن آنها می‌تواند به بیان واقعیت فیزیک مسئله کمک قابل توجهی کند. در این مقاله عوامل بروز خطا در داده‌های برداشت شده از بویه‌های موج‌نگار بررسی شد و از روش ضریب داده پرت محلی که یکی از قدرتمندترین روش‌های یادگیری ماشین جهت تشخیص ناهنجاری در داده‌ها است، جهت شناسایی داده‌های مشکوک به خطا در یک نمونه واقعی از داده‌های برداشت شده از بویه‌های موج‌نگار استفاده شد. مهمترین نتایج این تحقیق به شرح زیر است:

- بررسی‌ها حاکی از عملکرد مناسب روش ضریب داده پرت محلی جهت شناسایی داده‌های مشکوک به خطا در نمونه دیتای موج مورد بررسی است.
- در نمونه‌های مورد بررسی روش ضریب داده پرت محلی به ازای توابع اقلیدسی، اقلیدسی استاندارد شده، بلاک شهری، مینکوسکی، چبیشف و ماهالانوبیس نتایج کاملاً مشابهی ارائه کردند. بنابراین می‌توان به دلخواه از هر یک از این توابع استفاده کرد. اما توابع کسینوس، همبستگی، اسپیرمَن، هَمینگ و جاکارد نتوانسته‌اند با موفقیت اجرا شوند. بنابراین بکارگیری این توابع با روش ضریب داده پرت محلی پیشنهاد نمی‌شود.
- داده‌های علامت‌گذاری شده توسط روش همواره بیانگر وجود خطا در داده‌ها نیست و چه بسا ممکن است به دلیل شرایط خاص مسئله ایجاد شده باشند. بنابراین پس از شناسایی این داده‌ها، باید در خصوص علت

وقوع آنها تحقیق کرده و در صـــورتی که خطا باشـــند نسبت به حذف و یا تصحیح آنها اقدام کرد.

## ۶- مراجع


[۱] واقفی، محمد، قدسیان، مسعود، و ادیب، آرش. "نگرشی بر خطاهای موجود در برداشت داده‌های آزمایشگاهی"، نهمین کنفرانس هیدرولیک ایران، دانشگاه تربیت مدرس، آبان ماه ۱۳۸۹.

[۲] کتابداری، محمد جواد، و شهربیجاری، علی نظری. "مدل کامپیوتری جامع جهت آنالیز دیتاهای خام ناشی از بویه‌های موج نگار"، هشتمین همایش صنایع دریایی، بوشهر، ۹ لغایت ۱۱ آبانماه ۱۳۸۵.

[3] http://marinedata.pmo.ir

[4] Bolton, R., and Hand, D. "Unsupervised Profiling Methods for Fraud Detection", In Credit Scoring and Credit Control VII, 1999.

[5] Eskin, E., "Outlier Detection Over Noisy Data using Learned Probability Distributions", In Proceedings of the 7th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., pp.255–262, 2000.

[6] Aggarwal, C., "On Abnormality Detection in Spuriously Populated Data Streams", In Proceedings of 5th SIAM Data Mining, pp.80–91, 2005.

[7] Manson, G. "Identifying Damage Sensitive, Environment Insensitive Features for Damage Detection", In Proceedings of the IES Conference, Swansea, UK, pp. 1-5, 2002.

[8] Hollier, G., Austin, J. "Novelty Detection for Strain-Gauge Degradation using Maximally Correlated Components", In Proceedings of the European Symposium on Artificial Neural Networks, pp.262–539, 2002.

[۹] محمودی، کیومرث، و سایبانی، مصباح. "سلامت‌سنجی سیستم‌های دریایی به روش تشخیص ناهنجاری در یادگیری ماشین"، اولین همایش ملی فناوری‌های نوین دریایی (MMT2013)، مازندران، دانشگاه علوم دریایی امام خمینی (ره)، ۱۶ الی ۱۷ تیر ۱۳۹۲.

[10] Augusteijn, M., Folkert, B. "Neural Network Classification and Novelty Detection", International Journal on Remote Sensing, Vol. 23, No. 14, pp. 2891–2902, 2002.

[11] Theiler, J., and Cai, D. M. "Re-sampling Approach for Outlier Detection in Multispectral Images", In Proceedings of SPIE, Vol.5093, pp.230-240, 2003.

[12] Spence, C., Parra, L., and Sajda, P. "Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model", In Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, IEEE Computer Society, Washington, DC, USA, Vol. 3, 2001.

[13] Pokrajac, D., Lazarevic, A., and Latecki, L. J. "Incremental Local Outlier Detection for Data Streams", In Proceedings of IEEE Symposium on Computational Intelligence and Data Mining, pp.1-11, 2007.

[14] Yankov, D., Keogh, E., and Rebbapragada, U. "Disk Aware Discord Discovery: Finding Unusual Time Series in Terabyte Sized Datasets", Knowledge and Information Systems, Vol.17, No.2, pp.241–262, 2008.

[15] Solberg, H. E., and Lahti, A. "Detection of Outliers in Reference Distributions: Performance of Horn's algorithm", Clinical Chemistry, Vol.51, No.12, pp.2326–2332, 2005.

[16] Suzuki, E., Watanabe, T., Yokoi, H., and Takabayashi, K. "Detecting Interesting Exceptions From Medical Test Data With Visual Summarization", In Proceedings of the 3rd IEEE International Conference on Data Mining, pp.315–322, 2003.

[17] Chen, D., Odobez, J. M., and Bourlard, H. "Text Detection and Recognition in Images and Video Frames", Pattern Recognition, Vol.37, No.3, pp.595-608, March 2004.

[18] Breuning, M., Kriegel, H-P., Ng, R., and Sander, J., "LOF: Identifying Density Based Local Outliers", In Proc. of ACM SIGMOD International Conference on Management of Data (SIGMOD'00), Dallas, Texas, pp.93-104, 2000.

[19] Chen, M. C., Wang, R. and Chen, A. P., "An Empirical Study for the Detection of Corporate Financial Anomaly Using Outlier Mining Techniques", In ICCIT'07: Proc. of the International Conferenceon Convergence Information Technology, pp.612–617, 2007.

[20] Xi, J., "Outlier Detection Algorithms in Data Mining", Intelligent Information Technology Applications", Vol.1, pp.94–97, 2008.

[۲۱] محمودی، کیومرث، کتابداری، محمد جواد، و سایبانی، مصباح. "شناسایی نفوذ به شبکه‌های کامپیوتری سیستم‌های نظامی، به روش تشخیص ناهنجاری"، اولین همایش ملی فناوری‌های نوین دریایی، نوشهر، دانشگاه امام خمینی (ره)، ۱۲ تا ۱۳ شهریور ماه ۱۳۹۲.

[۲۲] محمودی، کیومرث، سایبانی، مصباح، و مرادی، عباس. "شـناسایی خـطاهـای موجود در برداشت داده‌های آزمایشگاهی، با استفاده از الگوریتم‌های تشخیص داده خطا"، پانزدهمین کنفرانس دینامیک شاره‌ها FD2013، بندرعباس، دانشگاه هرمزگان، ۲۷- ۲۹ آذر ۱۳۹۲.