دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گروه هوش مصنوعی

پایان نامه کارشناسی ارشد

شناسایی هوشمند زون‌های شکسته در مخازن کربناته با استفاده از چاه‌نمودارهای پتروفیزیکی

امیر کشاورزرضا

اساتید راهنما:

دکتر حمید حسن‌پور و دکتر بهزاد تخم‌چی

استاد مشاور:

محمد حسین خسروی

ماه و سال انتشار :

آذر ۱۳۹۳

دانشگاه صنعتی شاهرود

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گروه هوش مصنوعی

پایان نامه کارشناسی ارشد آقای امیر کشاورزرضا

تحت عنوان: شناسایی هوشمند زون‌های شکسته در مخازن کربناته با استفاده از چاه-نمودارهای پتروفیزیکی

در تاریخ ۹۳/۶/۲۵ توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی ارشد مورد ارزیابی و با درجه عالی مورد پذیرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
دکتر حمیدحسن‌پور		محمدحسین خسروی	
دکتر بهزاد تخم‌چی			

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
دکتر محمدرضا کرمی		مهندس محسن فرهادی	
دکتر وحید ابوالقاسمی			

تقدیم به پدر و مادر فداکارم اسوه صبر
و شکیبایی ناپذیری در زندگیم.

و

تقدیم به همسرم که محبت و وجودش
برایم امید و روشنی مسیر زندگی است.

نخست خداوند بزرگ و علیم را شاکرم که به من قدرت تفکر و نوشتن داد و تلاش را در وجودم برای یافتن پاسخ به معماهایی هر چند کوچک از علم قرار داد. سپاس که توانستم با دانش اندک خود گامی در مسیر علم بردارم.

بر خود فرض می‌دانم که از اساتید راهنمای فرهیخته و دانشمند خویش آقایان دکتر حمید حسن پور و دکتر بهزاد تخم چی کمال تشکر و قدردانی را بجا آورم. در برابر لطف و توجه بزرگوارانه شان زانوی ادب بر زمین مینهم.

و نیز از خانواده‌ام و همه دوستانم که مرا در این کار یاری نمودند کمال تشکر را بجا می‌آورم. از دوستان خویش آقایان مهندس سید حجت مقدسی و مهندس محمود سیف الهی به طور خاص کمال تشکر را دارم.

تعهد نامه

اینجانب **امیر کشاورزرضا** دانشجوی دوره کارشناسی ارشد رشته **کامپیوتر** گرایش **هوش مصنوعی** دانشکده **مهندسی کامپیوتر و فناوری اطلاعات** دانشگاه شاهرود نویسنده پایان‌نامه با عنوان:

شناسایی هوشمند ویژگی‌های شکستگی‌ها بوسیله‌ی چاه‌نمودارهای پتروفیزیکی

تحت راهنمایی آقایان **دکتر حمید حسن‌پور** و **دکتر بهزاد تخم‌چی** متعهد می شوم.

- *تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.*
- *در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.*
- *مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.*
- *کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه شاهرود » و یا «Shahrood University» به چاپ خواهد رسید.*
- *حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.*
- *در کلیه مراحل انجام این پایان نامه، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.*
- *در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.*

تاریخ ***امضای دانشجو***

مالکیت نتایج و حق نشر

- *کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه های رایانه ای، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه شاهرود می باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.*
- *استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.*

چکیده

حدود ۵۰ درصد مخازن نفتی در دنیا شکافدارانند؛ که البته این نسبت در ایران به بیش از ۹۰ درصد می‌رسد. این در حالیست که تعیین دقیق شکستگی‌های یک میدان در حال حاضر بسیار پر هزینه بوده و مستلزم توصیف و مستندسازی گام به گام تعداد زیادی چاه‌نمودار تصویری توسط متخصصین خبره می‌باشد، بعلاوه‌ی اینکه در حال حاضر و به خصوص در ایران موانع جدی در مقابل لاگ‌برداری تصویری وجود دارد که نیاز به استفاده از سایر داده‌ها بمنظور شناسائی شکستگی‌ها را اجتناب ناپذیر کرده است.

تا کنون مطالعات متنوعی در زمینه تفکیک شکستگی‌های طبیعی باز و بسته صورت پذیرفته است که هنوز به هدف نهایی و کارآمد برای استفاده در صنعت نرسیده است. در این تحقیق سعی شده است با استفاده از داده‌های پتروفیزیکی به این تشخیص دست یابیم که آیا یک ناحیه از چاه دارای شکستگی باز، شکستگی بسته و یا فاقد شکستگی می‌باشد.

در روش پیشنهادی به جای استفاده از داده‌های خام با استفاده از تکنیک پنجره‌گذاری و جایگزین نمودن داده‌ها با واریانس پنجره هر سیگنال داده را به یک سیگنال ساده‌تر تبدیل نموده و از آن برای داده‌کاوی استفاده نمودیم. در دسته بندی داده‌ها از دو الگوریتم دسته‌بندی استفاده گردیده که در نهایت نتایج این دو الگوریتم با استفاده از روش رأی‌گیری اکثریت با هم ترکیب شده‌اند.

دو الگوریتم مورد استفاده در این تحقیق الگوریتم Naïve Bayes و Random Tree می‌باشند که در مقایسه با سایر الگوریتم‌های آزمایش شده در این تحقیق نتایج بهتری را از خود نشان داده‌اند.

این مطالعه بر روی چاه‌های میدان نفتی گچساران صورت پذیرفت. داده‌هائی که در اختیار این پروژه قرار گرفت دارای محدودیت‌هائی بود و بنابراین محدودیت‌ها قسمتهایی از آزمایشات بر روی داده‌های ۷ چاه با ۷ چاه‌نمودار انجام شده و قسمتهایی نیز بر روی داده‌های ۴ چاه با ۱۱ چاه‌نمودار.

گرچه هدف اصلی و رویکرد مطالعات در این تحقیق دسته‌بندی داده‌ها به سه کلاس شکستگی باز، شکستگی بسته و یا فاقد شکستگی بوده است ولی بمنظور آنکه دست اندرکاران صنعت نفت بتوانند مقایسه‌ای با نتایج تحقیقات خود داشته باشند در پایان بخشی نیز به دسته‌بندی دو کلاس دارای شکستگی و فاقد شکستگی، اختصاص یافته است.

در این تحقیق ایجاد توازن در میان صحت دسته‌بندی سه کلاس از اهمیت فوق‌العاده‌ای برخوردار است و این با طبیعت شکستگی‌ها سازگاری ندارد چرا که این داده‌ها دارای توازن نبوده و کلاس فاقد شکستگی ، با تفاوت بسیار زیاد کلاس غالب می‌باشد، لذا با درنظر گرفتن این اصل ما مجبور هستیم از نتایجی که دارای توازن کمتر می‌باشند حتی اگر از صحت عملکرد خوبی برخوردار باشند، صرف نظر کنیم. در پایان، برای دسته بندی دو کلاسه‌ی قابل تعمیم بهترین نتیجه در یکی از چاه‌ها به دقت

۷۷,۵ درصد و برای دسته‌بندی سه کلاسه‌ی قابل تعمیم برای یکی از چاه‌ها صحت عملکرد به بیش از ۸۰,۶ درصد رسید. این در حالیست که که میانگین دقت روش پیشنهادی بیش از ۷۰,۱ درصد را در دسته‌بندی سه کلاسه‌ی قابل تعمیم و نیز بالاتر از ۶۸,۶ درصد را در دسته‌بندی دو کلاسه‌ی قابل تعمیم از خود نشان می‌دهد.

کلمات کلیدی:

داده‌کاوی، شکستگی، شکستگی باز، شکستگی بسته، دسته‌بندی، الگوریتم Naïve Bayes، الگوریتم Random Tree، Majority Voting، چاه‌نمودار پتروفیزیکی، چاه‌نمودار تصویری، پنجره‌گذاری.

مقالات استخراج شده

۱. کشاورزرضا امیر، حسن‌پور حمید، تخم‌چی بهزاد، ۱۳۹۳، شناسایی هوشمند زون‌های شکسته با استفاده از چاه‌نمودارهای پتروفیزیکی، اولین کنفرانس ملی مهندسی اکتشاف منابع زیرزمینی، شاهرود، ایران.

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول

مقدمه و کلیات

۱-۱- مقدمه

در سرتاسر جهان مخازن شکافدار حدود ۵۰ درصد از ذخائر هیدروکربنی جهان را به خود اختصاص داده‌اند که البته این مقدار در ایران بیشتر نیز می‌باشد. مطالعات نشان داده که شکستگی‌ها تاثیر به سزایی در میزان تراوایی مخازن نفتی دارد. بعنوان مثال مستند به مطالعات انجام شده، اگر یک شکستگی منفرد با بازشدگی یک میلیمتر در سنگ مخزن باشد و توسط چاهی قطع شود، یک تراوایی خوب و مناسب را ایجاد می‌کند، که تولید نفتی در حدود ۱۰۰۰ متر مکعب در روز را باعث می‌شود [۱].

تا کنون تلاش‌های زیادی برای شناسایی شکستگی‌ها صورت گرفته است که از آن میان می‌توان به استفاده از داده‌های لرزه نگاری [۲]، داده‌های پتروفیزیکی [۳]، تست چاه [۴]، بررسی نحوه از دست رفتن گل حفاری و تفسیر مغزه اشاره نمود [۵]. مقالاتی نیز در خصوص شناسایی شکستگی‌ها از جمله [۶] و چگالی شکستگی‌ها همانند [۷] از روی داده‌های پتروفیزیکی منتشر شده که می‌تواند بعنوان مرجع خوبی برای تحقیقات در این زمینه باشد.

البته با ظهور چاه‌نمودارهای[۱] تصویری و عمق‌سنج در اواسط دهه ۱۹۸۰ شناسایی ویژگی‌های شکستگی‌ها بسیار ساده‌تر شد [۸] اما متاسفانه در کشور ایران تعداد بسیار زیادی چاه وجود دارند که امکان تهیه چاه‌نمودارهای تصویری از آنها غیر ممکن است [۶] و این در حالیست که اطلاعات پتروفیزیک آنها در دست است و این خود می‌تواند انگیزه بزرگی برای انجام این تحقیق باشد.

۱-۲- ضرورت انجام کار

مدل‌سازی شکستگی مستلزم داشتن داده از شکستگی‌ها است. داده‌های شکستگی نیز مستخرج از لاگهای تصویری است که کمتر در دسترس‌اند. لذا شناسایی شکستگی‌ها از ابزار دیگری همچون

[۱] Well logs

لاگهای پتروفیزیکی مفید خواهد بود. از طرف دیگر برای افزایش هر چه بیشتر صرفه اقتصادی در امر توسعه یک میدان، تعیین هر چه دقیق‌تر شکستگی‌ها بعنوان مهمترین عامل کنترل کننده جریان سیال، در نهایت می‌تواند منجر به مدل‌سازی هر چه بهتر ساختار فیزیکی و دینامیکی میدان گردد.

در سالیان گذشته رغبت به مطالعه شکستگی‌های طبیعی در سازندهای سطحی و زیر سطحی به طرز حیرت‌آوری افزایش یافته که این امر تا حد زیادی بر آمده از دانش کلان صنعتی در باب تاثیر شکستگی‌ها بر جریان سیال در زیر زمین می‌باشد [۱]. از سوئی دیگر باید به این نکته توجه داشته باشیم که عوامل مؤثر بر ساختار شکستگی‌های طبیعی به قدری زیاد هستند که آگیلریا[۱] در [۹] بیان داشته که به منظور بهره‌برداری از مخازن دارای شکستگی طبیعی، هر مخزن باید بعنوان یک پروژه‌ی تحقیقاتی مورد مطالعه قرار گیرد.

تعیین دقیق شکستگی‌های یک میدان در حال حاضر بسیار پر هزینه بوده و مستلزم توصیف و مستندسازی گام به گام تعداد زیادی چاه‌نمودار تصویری توسط متخصصین خبره می‌باشد. تهیه چاه نمودار تصویری بسیار هزینه‌بر بوده و بنا به دلایل مختلف هم اکنون تهیه آن برای هزاران چاه حفر شده در ایران غیر ممکن می‌باشد تا حدی که در حال حاضر تنها در کمتر از ده درصد چاه‌های کشور این چاه‌نمودارها رانده شده‌اند، لذا در این تحقیق سعی می‌شود با استفاده از داده‌هایی که دستیابی به آنها با هزینه ای به مراتب کمتر و با در صد بسیار بالاتر قابل حصول می‌باشد اقدام به تعیین ویژگی-های شکستگی‌ها نمائیم. این داده‌ها در اصطلاح چاه‌نمودارهای پتروفیزیکی نامیده می‌شوند که در فصل سوم به تشریح مورد بحث قرار خواهد گرفت. منظور از چاه‌نمودارهای پتروفیزیکی کلیه نمودارهای برداشت شده از چاه‌های حفر شده در مخازن نفتی توسط سوندهای مخصوص است. از این جمله می‌توان به چاه‌نمودارهای سونیک، الکتریکی، چگالی، و نوترون اشاره کرد.

[۱] Aguilera

تا این زمان تحقیقات کافی در خصوص شناسایی شکستگی‌ها و ویژگی‌های آنها صورت نگرفته است، به نحوی که نیاز به این مسئله در صنعت کاملا مشخص می‌باشد. برای کافی نبودن این تحقیقات می‌توان دو دلیل را ذکر نمود، یکی اینکه نیاز به آن محدود به مناطق خاصی از جهان می‌باشد و دوم اینکه رفتار شکستگی‌ها بسیار پیچیده بوده و مطالعه و تحقیق پیرامون این موضوع ذاتا کاری دشوار و حتی برای برخی غیر ممکن می‌نماید.

۱-۳- مروری برکارهای گذشته

تا کنون از روش‌های متنوعی برای شناسایی شکستگی‌ها و ویژگی‌های آنها استفاده شده است که اگر بخواهیم آنها را نام ببریم می‌توان به استفاده از مقاطع لرزه‌ای، چاه نمودارها، آزمایش چاه، هرزروی گل و توصیف مغزه‌ها اشاره نمود. این روش‌ها عموما با محدودیت‌هایی جدی همراه بوده‌اند که استفاده از آنها را در صنعت دشوار و یا غیر ممکن می‌نماید. علت نارسایی بیشتر این روش‌ها عدم قدرت تفکیک مناسب و در نتیجه بهره نامناسب برای شناسایی شکستگی‌ها می‌باشد، البته به جز توصیف مغزه‌ها که بعلت ضریب بازیافت پائین در زون‌های شکسته و نیز اصولا موجود نبودن به دلیل هزینه بالا استفاده از آن نیز روش مناسبی نخواهد بود. دسترسی به نتایج بسیاری از تحقیقات انجام شده در زمینه‌ی شناسائی شکستگی‌ها به سادگی میسر نمی‌باشد. در زیر سعی شده به مرور مستندات منتشر شده در سالهای اخیر پرداخته شود.

زولو[۱] و آونس[۲] [۱۰] در تحقیقی تحت عنوان محاسبات نرم[۱] ابتدا با استفاده از منطق فازی[۲] به بررسی رابطه‌ی میان داده‌های زمین‌شناسی و شکستگی‌ها پرداختند و در گام بعدی سعی نمودند که با استفاده از شبکه‌ی عصبی[۳] رابطه‌ای میان این داده‌ها و وجود شکستگی‌ها ایجاد نمایند.

---

[۱] Zellou
[۲] Ouenes

نعمتی و پزشک [۱۱] به بررسی تأثیر نوع سنگ و همچنین تخلخل و تراوایی بر روی فراوانی شکستگی‌ها با استفاده از تفسیر مغزه پرداخته‌اند و نتیجه‌گیری کرده‌اند که با افزایش دولومیت، فراوانی شکستگی‌های آسماری افزایش می‌یابد، این درحالیست که با افزایش سیلیس یا انیدریت، فراوانی شکستگی‌ها کاهش می‌یابد. ایشان همچنین مشاهده نموده‌اند که تغییرات میزان آهک، تخلخل و تراوایی، تأثیر چندانی بر فراوانی شکستگی‌های آسماری ندارد.

ونبرگ و همکاران[۴] [۱۲] مطالعاتی را در طاقدیس خویز انجام داده‌اند، ایشان ضخامت مکانیکی لایه‌ها و بافت آن‌ها را به عنوان دو عامل مهم کنترل کننده فراوانی شکستگی‌ها شناخته‌اند.

ساگی و همکاران[۵] [۱۳] برای اندازه‌گیری شکستگی‌های کوچک مقیاس و بررسی اتصال آن‌ها، استفاده از دو روش مجزا و برآورد نتایج را پیشنهاد می‌کنند. آنها در تحقیقات خود از دو روش تحلیل تصاویر دوبعدی استفاده نموده‌اند.

در یک پروژه‌ی داده‌کاوی محبی و همکاران [۱۴] از اعمال تبدیل ویولت بر چاه‌نمودارهای پتروفیزیکی و یا اطلاعات چاه استفاده نموده‌اند و توانسته‌اند پی به‌وجود شکستگی‌های آسماری در دو مخزن متفاوت ببرند. ایشان تغییرپذیری‌های نهفته در فرکانس‌های بالای چاه‌نمودارها را به‌عنوان معرف شکستگی تفسیر کرده‌اند.

استفنسون و همکاران[۶] [۱۵] در مطالعات خود بر روی چند طاقدیس زاگرس، رابطه زون‌های اصلی شکسته و چین‌خوردگی زاگرس را بررسی کرده‌اند. ایشان نتیجه‌گیری کرده‌اند که زون‌های اصلی شکسته، در راستای محور طاقدیس‌ها بوده، و اتفاقاً همین زون‌ها هستند که نقش اصلی را در جریان سیال در مخازن نفتی زاگرس بازی می‌کنند.

---

[۱] Soft Computing
[۲] Fuzzy Logic
[۳] Neural Network
[۴] *Wennberg et al.*
[۵] Sagi
[۶] *Stephenson et al.*

در یک پروژه‌ی داده‌کاوی دیگر تخم چی و همکاران [۶] با ترکیب آنالیز چاه‌نمودارهـای پتروفیزیکـی، کلاسه‌بندی و تلفیق اطلاعات روشی نوین جهت شناسایی زون‌های شکسته معرفـی کـرده‌انـد. در ایـن مطالعه از داده‌های هشت چاه سارند آسماری جهت معرفی و سنجش صحت روش استفاده شده اسـت. از روش تطابق انرژی جهت شناسایی ویولت مبناء بهینـه کـه در کـاهش نوفـه و تجزیـه پتروفیزیکـی کاربرد دارد، استفاده شده است. به کمک دسته‌بندی کننده‌های پارزن[۱] و بیزین داده‌هـای خـام بـدون نوفه در باندهای فرکانسی مختلف کلاسه‌بندی شده و نتایج نشان داده که باندهای فرکانسـی پـایین از چاه‌نمودارهای بدون نوفه، بهترین داده‌ها جهت شناسایی زون‌های شکسته بوده‌اند.

تخم چی و همکاران [۱۶] در مطالعه‌ای دیگر روشی نوین برای تشخیص زون شکستگی و خصوصـیات شکستگی‌ها به کمک چاه‌نمودار اشباع آب ارائه کردند. برای تشخیص زون شکستگی از تبدیل ویولـت، که ابزاری کارآمد جهت تشخیص تغییرات محلی در داده‌هـا مـی‌باشـد، اسـتفاده شـده اسـت. در ایـن مطالعه برای دستیابی به باندهای فرکانسی باریک‌تر، از ویولت بسته‌ای استفاده شده است. تجریه چـاه-نمودار اشباع آب به کمک ویولت نشان داده که بخـش اعظـم اطلاعـات در بانـدهای فرکانسـی پـایین نهفته است، بنابراین از بخش تقریب جهت تشخیص شکستگی و از حجم شیل (یـا چـاه‌نمـودار گامـا) برای حدف خطای تخمین و تعیین زون ابهام استفاده شده است. در نهایت رابطه‌ای خطی میان انرژی بخش تقریب چاه‌نمودار اشباع آب و تراکم شکستگی‌ها به دست آمده است که منجر به تخمـین تعـداد شکستگی در زون شکستگی می‌گردد

در مطالعه‌ای دیگر، سروش و همکاران [۱۷] بـه کمـک ترکیـب روش‌هـای بیـزین، ویولـت و ترکیـب اطلاعات به شناسایی بازشدگی در چاه‌های کربناته پرداخته‌اند.

در مطالعه‌ای مشابه سروش و همکاران [۱۸] با استفاده از روشـی چنـد متغیـره زون‌هـای ریزشـی در مخازن گازی شیلی را با به‌کارگیری داده‌های چاه‌نمودارهای پتروفیزیکی، وزن گـل و داده‌هـای تـنش

[۱] Parzen

روباره مشخص کرده‌اند. در این روش از تکنیک‌های پردازش داده مانند کلاسه بندی بیزین، تجزیه ویولت و تلفیق اطلاعات جهت تعیین محدوده‌های ریزش استفاده شده است.

ژانگ و همکاران در مطالعاتی از این دست [۱۹]، از ایده نوین ترکیب تبدیل ویولت و منحنی‌های تفاضلی چاه‌نمودارهای مرسوم جهت شناسایی زون‌های شکسته بهره برده‌اند. در ابتدا چاه‌نمودارهای مرسوم به کمک ویولت‌های مبناء تجزیه و سیگنال‌های تجزیه شده با شکستگی‌های حاصل از چاه-نمودارهای تصویری جهت انتخاب ویولت مبناء متناسب با شکستگی‌ها مقایسه شده است. سپس با ترکیب سیگنال‌های تجزیه شده با ویولت مبناء متناسب با شکستگی‌ها و منحنی‌های تفاضلی چاه-نمودارهای مرسوم منحنی شاخص شکستگی تشکیل شده است.

در سال‌های اخیر استفاده از الگوریتم‌های دسته‌بندی و مفاهیم داده‌کاوی در صنعت نفت جایگاه خوبی را داشته است. در یکی از این نوع تحقیقات مسعودی و همکاران [۲۰] موفق شدند با استفاده از چهار الگوریتم دسته‌بندی بر روی داده‌های پتروفیزیکی و آزمون چاه و ترکیب نتایج این چهار الگوریتم، میدان نفتی سراوان را به هفت ناحیه تقسیم کرده و در هر یک از این نواحی تعیین کنند که استخراج منجر به تولید آب خواهد شد و یا نفت و یا بدون تولید خواهد بود.

مسعودی و همکاران در تحقیقی مشابه [۲۱] در مطالعه‌ای بر چهار چاه نفت در جنوب ایران با بکار گیری دو الگوریتم دسته‌بندی بیزین و ترکیب نتایج آن‌ها توسط تئوری فازی، به تعیین نواحی دارای نفت پرداختند. آن‌ها در این تحقیق استفاده از تئوری فازی در ترکیب نتایج را بر روی بالا بردن صحت عملکرد و نیز بالا بردن توان تعمیم الگوریتم مؤثر دانسته‌اند.

زروکی و همکاران[۱] [۲۲] برای برآورد میزان تخلخل شکستگی‌ها در بیشتر چاه‌های منطقه‌ی هاسی مساود[۲] در کشور الجزایر با مشکل فقدان وجود یکی از لاگ‌های مورد نیاز یعنی زمان از دست رفتن گل حفاری مواجه بودند، که وجود این لاگ برای تعیین میزان تخلخل مورد نیاز کارشناسان صنعت

[۱] Zerrouki

[۲] Hassi Messaoud

نفت می‌باشد. راه‌حل آن‌ها برای برون رفت از این مشکل این بود که سایر لاگ‌ها را بعنوان ورودی یک شبکه عصبی مصنوعی به خدمت گرفته و موفق به برآورد میزان تخلخل گردیدند. آن‌ها با استفاده از رتبه‌بندی فازی توانستند اثبات کنند که در داده‌های مورد مطالعه قادر به حذف کردن لاگ دیگری از میان ورودی‌های خود نبودند.

مسعودی و همکاران [۲۳] در آخرین تحقیق ارائه شده‌ی خود با انتخاب و تعیین مشخصه‌های مؤثر در تشخیص شکستگی‌ها، تراوائی و سایر ویژگی‌های سنگ پرداختند. آنها برای این کار از دو الگوریتم شبکه‌ی بیزین و K۲ استفاده نمودند و در نتیجه‌گیری خود چگالی بالک[۱] را بعنوان مؤثرترین مشخصه‌ی تعیین کننده‌ی ویژگی‌های مخزن بیان نموده‌اند. چگالی بالک عبارت است از جرم فاز جامد بر حجمی که اشغال می‌کند.

ملّاجان و همکاران [۲۴] مطالعه‌ای برای تعیین نواحی دارای تخلخل حفره‌ای[۲] در چهار چاه در سازند سروک واقع در غرب ایران انجام داده‌اند. در این تحقیق از دو الگوریتم دسته‌بندی کننده‌ی بیزین و پارزن و اعمال آنها بر روی چهار چاه‌نمودار پتروفیزیکی GR, NPHI, RHOB, DT استفاده شده و در آخر الگوریتم پارزن با دقت میانگین ۷۷,۷ درصد با بهینه سازی پارامترها در هر چاه و با دقت میانگین ۶۷,۳ درصد در بهینه سازی قابل تعمیم عملکرد مناسب‌تری را داشته است.

در یک تحقیق مشابه عسگری‌نژاد و همکاران [۲۵] با استفاده از همان چاه‌نمودارهای پتروفیزیکی باز هم به تعیین نواحی دارای تخلخل حفره‌ای پرداختند، با این تفاوت که این بار از ترکیب موجک و پارزن برای دسته‌بندی نواحی چاه استفاده نمودند. آن‌ها نتایج بدست آمده را با نتایج حاصل از تفسیر مغزه مقایسه نمودند و کار خود را موفق دانسته‌اند.

همانطور که پیش‌تر نیز اشاره کردیم، با توجه به قدرت تفکیک خوب چاه نمودارهای تصویری می‌توان از آنها برای برطرف نمودن این مسئله بهره گرفت، این روش در اواسط دهه ۱۹۸۰ میلادی ارائه گردید

[۱] Bulk density

[۲] Vuggy Porosity

و هم اکنون بیش از یک دهه از کاربرد صنعتی آن نمی‌گذرد[۸]. بنابراین دو مشکل اساسی وجود دارد که استفاده از آن را غیر ممکن می‌سازد، اول آنکه بیشتر چاه‌های کشور قبل از ورود این تکنولوژی به صنعت حفر شده‌اند و از اینرو امکان راندن این چاه‌نمودارها در آنها غیر ممکن است و دوم اینکه دانش فنی برداشت و تفسیر این چاه‌نمودارها تنها در انحصار چند شرکت بین‌المللی است که این مسئله استفاده از آن در حفاری‌های اخیر را با محدودیت‌های جدی مواجه می‌کند [۲۶].

۱-۴- ساختار پایان نامه

در این پایان نامه روش تشخیص ویژگیهای شکستگیهای زمین از روی داده‌های پتروفیزیکی بمنظور استفاده در تحلیل ساختار چاه‌های نفت بیان شده است. ساختار و محتوای پایان نامه شامل هفت فصل است. چکیده‌ای از فصل‌های آینده در ادامه بیان شده است.

فصل دوم- شکستگی‌ها و لاگ‌های چاه‌های نفت: در این فصل در مورد مفاهیم مربوط به شکستگی‌ها و چاه‌نمودارهای پتروفیزیکی و تصویری به نحوی مورد بررسی قرار گرفته‌اند که خواننده بتواند درک خوبی از فضای مورد تحقیق بدست آورد. متخصصان صنعت نفت می‌توانند از این فصل صرف نظر نمایند.

فصل سوم- مفاهیم داده‌کاوی: داده‌کاوی یکی از زمینه‌های اصلی هوش مصنوعی بوده و بصورت بسیار مؤثر در اختیار بسیاری از علوم قرار گرفته. با توجه به گستردگی مباحث داده‌کاوی در این فصل تنها به تشریح روش‌ها و مفاهیمی پرداخته شده است که مورد نیاز روش پیشنهادی پروژه‌ی پیش رو می-باشد.

فصل چهارم- ارائه روش پیشنهادی: در ابتدای این فصل به اختصار به معرفی مفاهیم اولیه به کار گرفته شده در این تحقیق از جمله روش‌های گوناگون استخراج ویژگی را بیان کرده و با ترکیب آنها راهکار ارائه شده برای تشخیص ویژگی‌های شکستگی‌ها را بیان می‌نمائیم.

فصل پنجم- داده‌های تحقیق و میدان مورد مطالعه: تشریح دقیق داده‌های تحقیق در این فصل به درک هر چه بهتر فضای مسئله و در نتیجه برنامه ریزی بهتر برای گام‌های بعدی کمک خواهد نمود.

فصل ششم- پیاده‌سازی و ارزیابی نتایج: در این فصل نحوه پیاده‌سازی روش پیشنهادی بیان شده و خروجی بدست آمده به صورت مبسوط ارائه می‌گردد. این فصل در بردارنده مهمترین اطلاعات این تحقیق می‌باشد.

فصل هفتم- نتیجه‌گیری و پیشنهادات: نتیجه‌گیری و پیشنهادات حاصل از این تحقیق بیان می‌شود. با توجه به نتایج به‌دست آمده از الگوریتم پیشنهادی، نقاط قوت و ضعف الگوریتم بیان می‌شود. برای بهبود و توسعه الگوریتم پیشنهادی و نیز کارهای آینده پیشنهادتی در پایان بیان می‌شود.

فصل دوم

شکستگی ها

و لاک های چاه های نفت

۲-۱- مقدمه

لاگ برداری از دیواره چاه، به معنای اندازه‌گیری خواص فیزیکی سنگ‌های دیواره چاه با روش‌های مختلف به منظور بررسی ویژگی‌ها و پدیده‌های موجود در دیواره و تبدیل این سیگنال‌ها به تصاویر مجازی، به منظور به تصویر کشیدن پدیده ها و ویژگی‌ها می‌باشد [۲۷].

هر نوع گسیختگی یا جدایش فیزیکی سنگ که ناشی از فزونی تنش‌ها بر مقاومت سنگ باشد که ممکن است به دلیل مکانیزم‌های طبیعی، عوامل حفاری و... در سنگ صورت گیرد شکستگی[۱] نامیده می‌شود [۲۸].

شکستگی را می‌توان به عنوان مهمترین پارامتر مؤثر در بهره‌برداری و اکتشاف میادین مواد کربناته دانست. این پدیده نقش بسیار مهمی را در مدل‌سازی چاه‌های کربناته دارد. شکستگی‌ها پارامتری تعیین کننده برای پایداری دیواره چاه و حرکت سیال در چاه‌های کربناته هستند.

۲-۲- هیدروکربن

در ساختار مولکولی هیدروکربن‌ها[۲] تنها عناصر کربن و هیدروژن شرکت دارند. هیدروکربن‌ها به دو دسته اصلی آلیفاتیک و آروماتیک تقسیم می‌شوند. هیدروکربن‌های آلیفاتیک خود به گروه‌های وسیع‌تری تقسیم می‌شوند: آلکانها، آلکنها، آلکینها و ترکیبات حلقوی مشابه. اکثر هیدرو کربن‌های موجود در زمین در نفت خام وجود دارند. در منابع مرتبط با مهندسی اکتشاف اغلب به مشتقات نفت و گاز هیدروکربن اطلاق می‌گردد.

---

[۱] Fracture

[۲] Hydrocarbon

۲-۳- توصیف یک مخزن

در مهندسی اکتشاف اولین هدف از پروژه‌های توصیف مخزن، گردآوری اطلاعات از منابع مختلف جهت شناسایی واحدهای جریانی[۱]، زون‌های دارای ویژگی‌های مشابه از نظر جریان سیال، و تعیین پیوستگی قائم و جانبی چنین زون‌هایی می‌باشد. سپس میانگینی از ویژگی‌های سیال و سنگ به منظور شبیه سازی رقومی مخزن به این زون‌ها نسبت داده می‌شود تا نهایتا بتوان رفتار مخزن را در طی تولید پیش‌بینی کرد. منابع متداولی که اطلاعات یک مخزن از آن بدست می‌آید شامل مغزه‌ها، نمودارها، آزمایش چاه، بررسی‌های لرزه‌ای و پیشینه‌ی تولید می‌باشد. توصیف یک مخزن برای کسب پارامترهای زیر ضروریست [۱]

- تعیین ظرفیت ذخیره
- تعیین هدایت هیدرولیکی
- تعیین پراکندگی نسبی تخلخل و تراوایی
- پیش‌بینی عملکرد یک مخزن
- برآورد مقدار تولید

۲-۳-۱- انواع مخازن هیدروکربنی

سنگ‌های مخزن بیشتر کربناته و یا ماسه سنگس هستند و به همین دلیل عمده‌ی بحث در سرتاسر دنیا بر مطالعه‌ی این مخازن می‌باشد [۱]. از اینرو در برخی از منابع مخازن هیدروکربنی را به سه دسته‌ی اصلی ماسه سنگی، کربناته و غیر معمول طبقه بندی نموده‌اند.

از آنجا که ما قصد داریم به شناسایی شکستگی‌های مخازن کربناته بپردازیم لذا از این پس فقط به بیان مفاهیم مرتبط با مخازن کربناته خواهیم پرداخت.

[۱] Flow units

۲-۳-۲- مخازن کربناته[1]

بیش از ۶۵ درصد هیدروکربن‌های خاورمیانه در مخازن کربناته هستند. سنگ‌های مخزن کربناته در میدان آغاجاری ایران روزانه بیش از ۱۰۰هزار متر مکعب نفت از حدود ۳۶ چاه تولید کرده‌اند. سنگ-های کربناته با خصوصیات ژنر، دیاژنر و پتروفیزیک خاص خود مشخص می‌شوند. منشا سنگ‌های کربناته، بر خلاف سنگ‌های تخریبی، در درون حوضه و یا نزدیک آن است. بیشتر آن‌ها منشا عالی داشته و محدوده‌ی وسیعی از اندازه ذرات را شامل می‌شود. در سنگ‌های کربناته تخلخل اولیه به سرعت از طریق پدیده‌های دیاژنزی کاهش می‌یابد.

۲-۳-۳- انواع تخلخل در کربنات‌ها

تخلخل موجود در کربنات‌ها بسیار متنوع است. یکی از طبقه‌بندی‌های بسیار مورد استفاده برای انواع تخلخل در سنگ‌های کربناته، طبقه‌بندی چوکت و پری است. در این طبقه بندی ۱۵ نوع تخلخل در ۳ گروه قرار می‌گیرند.(شکل ۲-۱)

گروه ۱. تخلخل‌های وابسته به فابریک شامل: بین دانه‌ای، درون دانه‌ای، بین بلوری، قالبی، فنسترال، چتری یا سایه‌بانی و رشدی

گروه ۲. تخلخل‌هایی که به فابریک ارتباطی ندارند شامل: شکستگی، حفره‌ای و کانالی، حفره‌ای خیلی بزرگ (cavern)

گروه ۳. تخلخل‌هایی که می‌توانند به فابریک وابسته باشند یا نباشند شامل: برشی(beccia)، گمانه‌ای (boring)، به هم ریختگی (burrowing)، و ترک‌های انقباضی (shrinkage).

زمان شکل‌گیری و منشا انواع این تخلخل‌ها در جدول ۲-۱ خلاصه شده است.

[1] Carbonate Reservoirs

جدول ۲-۱ زمان شکل‌گیری و منشأ انواع تخلخل در سنگ‌های کربناته

زمان شکل‌گیری	نوع تخلخل	منشأ
اولیه یا همزمان با رسوبگذاری	بین دانه‌ای درون دانه‌ای	رسوبی
ثانویه یا بعد از رسوبگذاری	بین بلوری	دولومیتی شدن
	قالبی و حفره‌ای	انحلال
	شکستگی	نیروهای ساختمانی

۲-۳-۴- مخازن شکسته

تمامی سنگ‌های رسوبی به جز تبخیری‌ها در شرایط خاص قابلیت شکنندگی را دارند. مقاومت کششی سنگ‌ها بسیار کمتر از مقاومت تراکمی آن‌هاست به همین دلیل در سنگ‌ها تحت تاثیر نیروی کششی درزه یا شکاف (Joint or Fracture) ایجاد می‌شود. درزه یا شکاف عبارت است از شکستگی-هایی که غالبا در سنگ‌ها موجود است و مهمترین مشخصه‌ی آن‌ها این است که حرکت نسبی به موازات صفحه‌ی شکستگی وجود ندارد. ابعاد شکستگی‌ها از چند سانتیمتر تا چندصد متر تغییر می-کند. سطح شکستگی‌ها در اکثر حالات یک سطح صاف و مستوی است. از آنجا که شکستگی‌ها از نظر هندسی به صورت صفحه‌ای می‌باشند، لذا برای مشخص کردن آنها باید شیب و امتداد آن مشخص گردد. سطح شکستگی ممکن است افقی، مایل و یا قائم باشد.

۲-۳-۵- نقش شکستگی‌ها در تخلخل و تراوایی

مطالعات شکستگی نشان داده است که اگر یک شکستگی منفرد به پهنای یک میلیمتر در سنگ مخزن باشد و توسط چاهی قطع شود یک تراوایی خوب و مناسب را ایجاد می‌کند، که تولید نفتی در حدود ۱۰۰۰ متر مکعب در روز را باعث می‌شود. تخلخل حاصل از شکستگی بدون ارتباط با فابریک سنگ و از نوع ثانویه است. ارتباط پهنای شکستگی با تراوایی به صورت زیر است:

$$K \qquad (۱-۲)$$

تراوایی بر حسب دارسی= k پهنای شکستگی بر حسب اینچ= W

۲-۴- پتروفیزیک

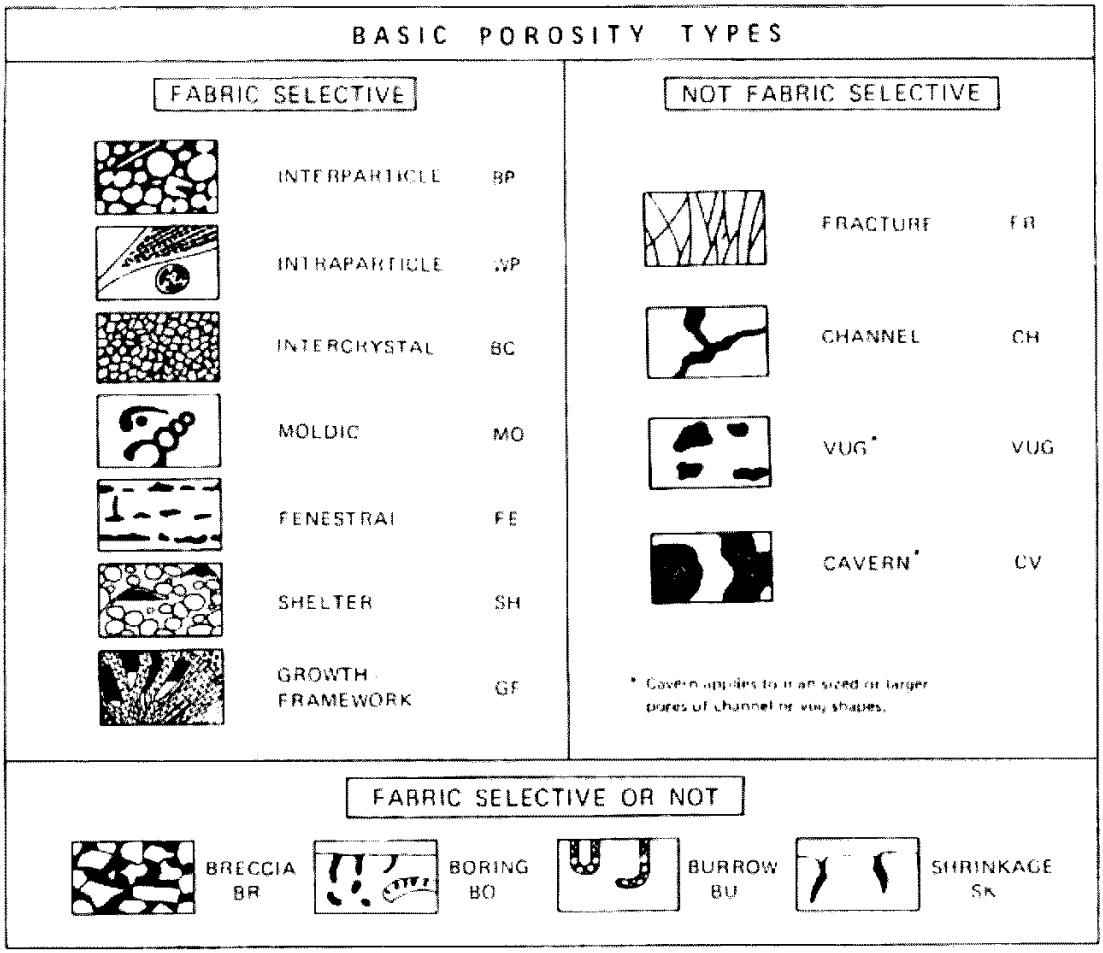
پتروفیزیک چیست: علمی است که به بررسی خواص فیزیکی مجموعه ی سنگ و سیال در مقابل تحریک های فیزیکی، امواج الکترومغناطیسی، الکتریکی، پرتوهای رادیواکتیو، امواج مکانیکی و ... می-پردازد.

داده های پتروفیزیکی: داده های به دست آمده از عملیات پتروفیزیک که شامل ثبت پاسخ به تحریکهای فیزیکی فوق است.

۲-۵- تصویربرداری از چاه

در تصویربرداری از دیواره چاه پس از اندازه‌گیری خواص فیزیکی سنگ‌های دیواره مخازن با روش‌های مختلف، به منظور بررسی ویژگی‌ها و پدیده‌های موجود در دیواره، این سیگنال‌ها به تصاویر مجازی که این پدیده‌ها و ویژگی‌ها را به تصویر می‌کشند تبدیل می‌شوند [۲۸].

به طور کلی با توجه به خاصیت مورد اندازه‌گیری سه تکنولوژی را می‌توان برای دستگاه‌های تصویر-بردار نام برد:



شکل ۲-۱ طبقه‌بندی انواع تخلخل در سنگ‌های کربناته توسط جوکت و پری [۱]

- **تصویر برداری الکتریکی**: در این روش از اختلاف مقاومت الکتریکی عوارض دیواره چاه برای به تصویر کشیدن آن‌ها استفاده می شود.
- **تصویر برداری صوتی**: تصویر به‌دست آمده از این روش، حاصل اختلاف تباین صوتی عوارض دیواره چاه است که به دو صورت زمان گذر امواج مافوق صوت و دامنه امواج مافوق صوت نمایش داده می شود.
- **تصویربرداری نوری**: در این روش با استفاده از یک منبع نوری و دوربین‌های ویدویی با لنزهای مخصوص به تصویربرداری از دیواره چاه می پردازند.

۲-۵-۱- روش تصویربرداری الکتریکی

این روش ابتدا در سال ۱۹۸۰ به عنوان تکنولوژی شیب سنجی[۱] معرفی شد. توسعه دستگاه‌های شیب سنج با تعداد بیشتر الکترود منجر به پیدایش و شکل‌گیری این دستگاه‌ها گردید. به منظور بهبود دستگاه‌های شیب سنج تعداد الکترودهای موجود در هر صفحه[۲] را افزایش دادند که این مسأله منجر به رسیدن به تصاویری با درجه تفکیک[۳] مناسب برای تشخیص شکستگی و لایه‌بندی‌ها در چاه‌ها شد که این دستگاه‌ها را تصویربردار الکتریکی نامیدند. این ابزارها با استفاده از یکسری الکترود نصب شده روی بالشتک‌ها قادر به اندازه‌گیری مقاومت میکرو یا هدایت میکرو در دیواره چاه هستند[۲۸]. دو کاربرد اصلی این دستگاه‌ها در شناسایی شکستگی‌ها و شناسایی و ارزیابی مخزن به‌وسیله‌ی تصاویر با درجه تفکیک بالا است. نسل‌های اول این دستگاه‌ها دارای چهار بالشتک بودند که هر بالشتک حاوی تعدادی الکترود برای تزریق جریان متناوب با فرکانس‌های خیلی پایین به دیواره چاه بودند.

در اوایل سال ۱۹۹۰ نیاز به تصاویر با پوشش بیشتر دیواره چاه منجر به ورود دستگاه‌های الکتریکی با شش بازو که قادر به پوشش ۸۰٪ دیواره چاه بودند شد. همچنین در دستگاه‌های چهار بازویی تعداد الکترودهای هر صفحه را افزایش دادند [۲۹].

با توجه به اینکه گل‌های حفاری که در حفاری و تصویربرداری استفاده می‌شوند در دو نوع گل-های رسانا (پایه آب) و نارسانا (پایه نفت) هستند به طور کلی می‌توان دستگاه‌های تصویربردار الکتریکی را به دو دسته زیر تقسیم کرد:

الف) دستگاه‌های تصویربردارالکتریکی در گل‌های رسانا (پایه آب)

[۱] Dipmeter
[۲] Pad
[۳] resolution

در این دستگاه‌ها تصویربرداری از طریق اندازه‌گیری مقاومت یا هدایت عوارض دیواره چاه در حالی که گل حفاری از نوع رسانا است صورت می‌گیرد. جریان تزریقی به سازند وارد دیواره شده و از گل حفاری عبور و به دیواره می‌رسد.

در زیر برخی از انواع این دستگاه‌ها معرفی می‌شود:

- **چاه‌نمودار FMS** [۱]

FMS نسل اول دستگاه‌های تصویربردار الکتریکی است که توسط شرکت شلومبرژه تولید شده است. انواع اولیه آن دارای دو بالشتک بوده که هر کدام دارای ۲۷ الکترود در چهار ردیف هستند و برای اندازه‌گیری مقاومت میکرو به‌کار می‌روند (شکل ۲-۲). این دستگاه در چاه ۸/۵ اینچی دارای پوشش ۲۰٪ است که جهت افزایش پوشش نمودارگیری، برداشت مجدد با چرخش دستگاه توصیه می‌شود. دستگاه چهار بالشتکی دارای ۱۶ الکترود در هر بالشتک است و جمعاً با اندازه‌گیری ۶۴ نمودار دارای پوشش ۴۰٪ در چاه‌های به قطر ۸/۵ اینچی خواهد بود. در طی نمودارگیری هر میکروالکترود یک جریان را به داخل سازند تزریق می‌کند. اندازه‌گیری شدت تغییرات جریان الکتریکی که تغییرات مقاومت میکرو سازند را منعکس می‌کند به تصاویری با شدت رنگ‌های مختلف خاکستری یا رنگی تبدیل می‌کند [۳۰].

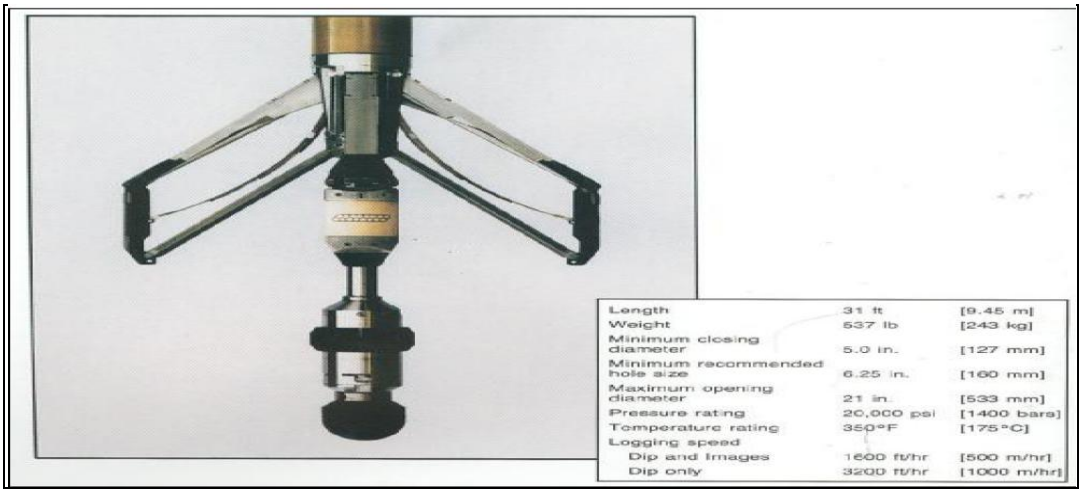
- **چاه‌نمودار FMI** [۲]

FMI نسل دوم دستگاه‌های تصویربردار الکتریکی درون‌چاهی است. در واقع یک زبانه به هر بازو نصب شده و در هر بالشتک و زبانه ۲۴ الکترود قرار دارد که باعث می‌شود، پوشش دیواره چاه تا ۸۰٪ افزایش یابد (شکل ۲-۳). قدرت تفکیک FMI، ۲/ اینچ (۵ میلی متر) بوده و توانایی تشخیص جزئیات تا ۵۰ میکرون را دارد. امروزه برای کاهش تعداد برداشت‌ها و افزایش پوشش دیواره چاه، FMI به FMS ترجیح داده می‌شود.

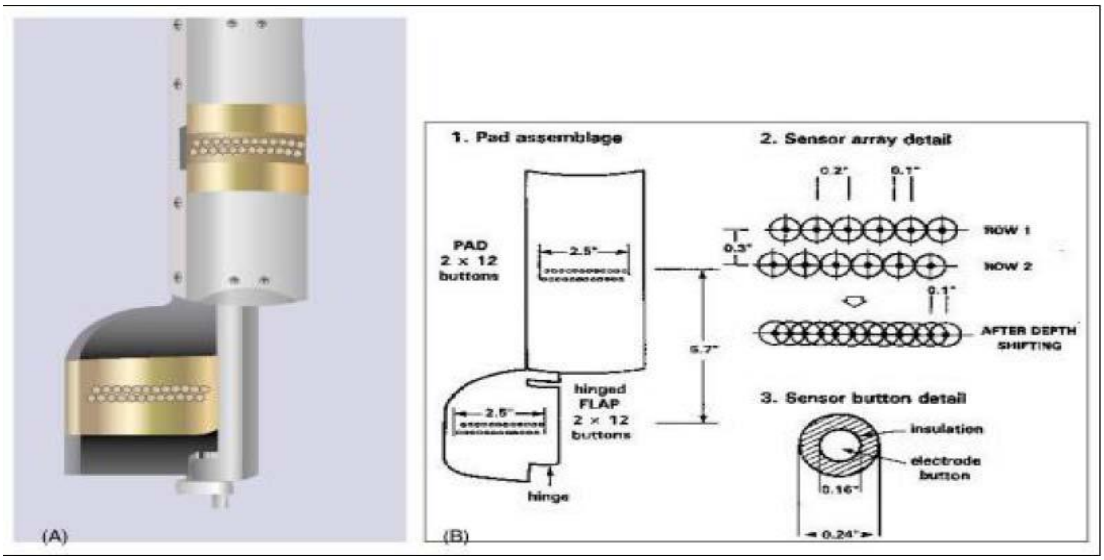
---

[۱] Formation Micro Scanner

[۲] Fullbore Formation Micro Imager



شکل ۲-۲ دستگاه تصویربردار FMS [۲۸]



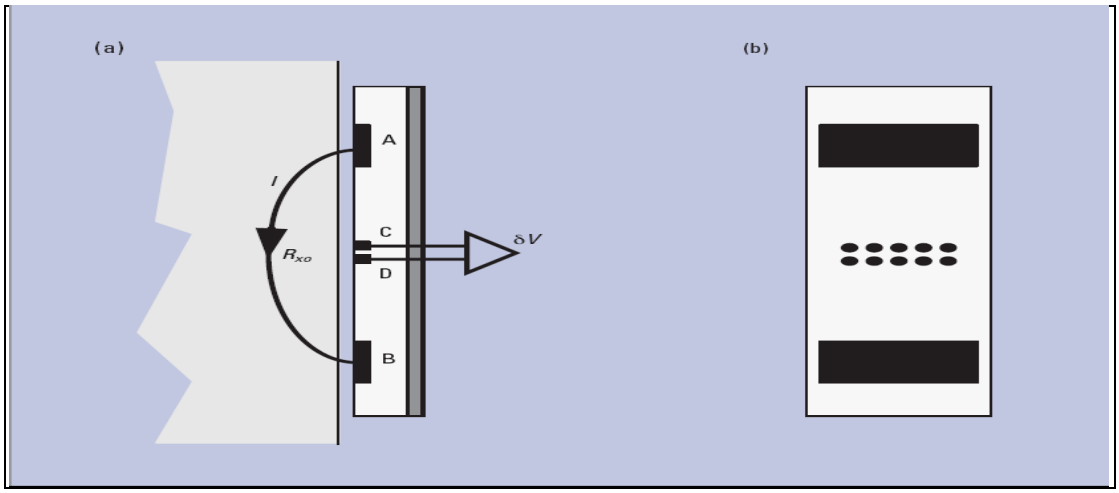
شکل ۲-۳ بالشتک و زبانه و مشخصات الکترودها در دستگاه تصویربردارFMI [۲۸]

ب) دستگاه‌های تصویربردارالکتریکی در گل‌های نارسانا (پایه نفت)

در این دستگاه‌ها تصویربرداری از طریق اندازه‌گیری مقاومت میکرو[۱] یا هدایت میکرو[۲] عوارض دیواره چاه در حالی که گل حفاری از نوع نارسانا است، صورت می‌گیرد. جریان تزریقی به سازند را به منظور امکان وارد شدن به دیواره با فرکانس‌های بالاتر تزریق می‌کنند تا بتواند از لایه عایق گل حفاری گذشته و وارد دیواره سازند شود ولی در عوض عمق نفوذ در دیواره کم می‌شود.

[۱] MicroResistivity
[۲] MicroConductivity

در برخی از چاه‌ها به‌خاطر خصوصیات سنگ مخزن، از گل پایه نفت استفاده می‌شود. این گل‌ها رسانایی الکتریکی پایینی داشته و امکان استفاده از ابزارهای تصویربردار الکتریکی معمولی مانند FMI و FMS وجود ندارد. بنابراین در این چاه‌ها ابزارهایی که قابلیت استفاده در گل‌های نارسانا را داشته باشند، به کار گرفته می‌شوند. افزایش استفاده از مواد نفتی و مصنوعی برای کاهش ریسک حفاری و از طرفی کیفیت تصویر برداشتی از سازند چالش‌هایی بودند که در تصویربردارهای الکتریکی معمولی وجود داشت [۲۸]. نمونه‌ای از این دستگاه‌ها OBMI است که در شکل (۲-۴) عملکرد تزریق جریان برای تصویربرداری نشان داده شده است.

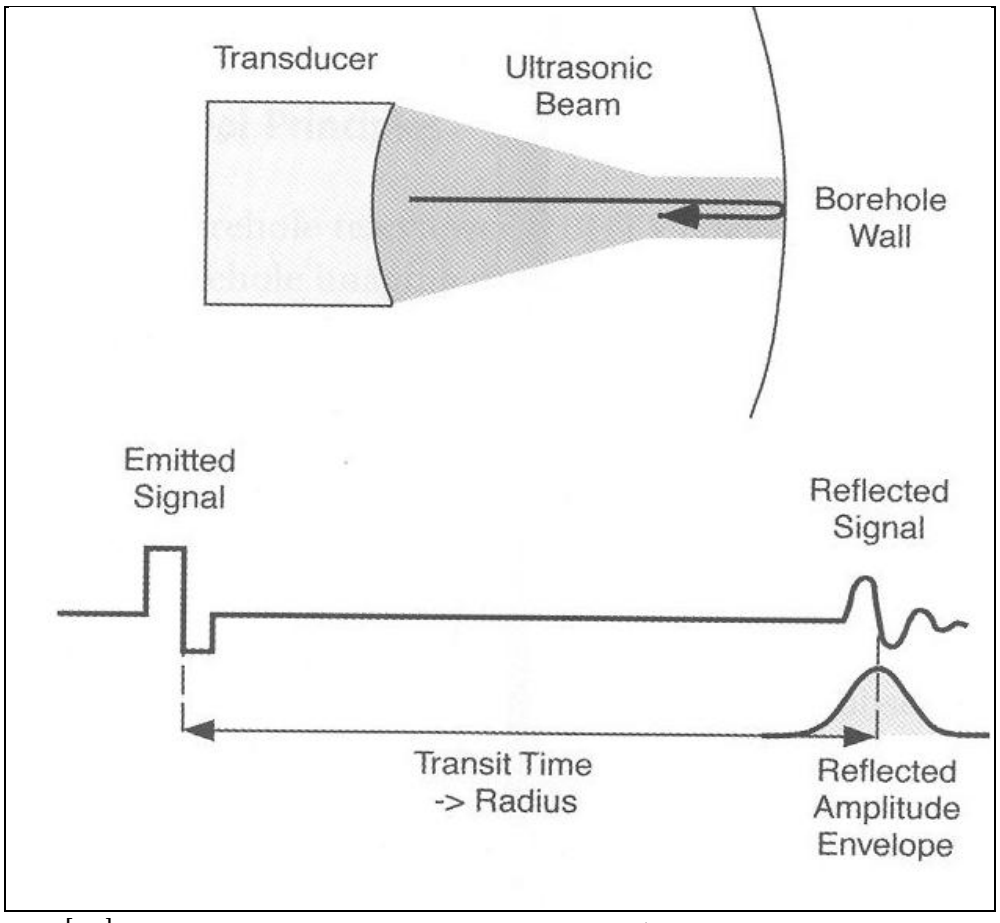


شکل ۲-۴ نحوه تزریق جریان به سازند در OBMI و آرایش الکترودها [۳۰]

۲-۵-۲- روش تصویربرداری صوتی

این تکنولوژی به‌وسیله شرکت Mobile در سال ۱۹۶۰ با نام نمایشگر چاه وارد صنعت نفت شد و اولین تکنولوژی مورد استفاده در صنعت نفت برای تصویربرداری بود. نمایشگرهای چاه قادر به اسکن دیواره چاه به‌وسیله امواج بازگشتی مافوق صوت از دیواره چاه هستند که تصاویر حاصل نتایج و ویژگی‌های جالبی را از دیواره چاه مانند شکستگی، لایه‌بندی و... فراهم می‌کند. این ابزارهای صوتی

وسایلی با مقطع دایره‌ای هستند که شامل یک فرستنده پیزوالکتریک[1] با سرعت چرخش rps ۱۲-۴ که در چاه می‌چرخد و امواجی را به‌صورت پالس در فرکانس‌های آلتراسونیک به دیواره چاه می‌فرستند [۲۷]. در این دستگاه‌ها فرستنده نقش گیرنده را نیز دارد و به اندازه‌گیری زمان رفت و برگشت و دامنه پالس منعکس شده از دیواره چاه در هر بار چرخش می‌پردازند. این تجهیزات در هر دو نوع گل-های رسانا و نارسانا حفاری قابل استفاده هستند [۲۹]. امواج منتشرشده از منبع پیزوالکتریک پس از برخورد به دیواره چاه منعکس شده و سپس به گیرنده که فرستنده، خود نقش آن را نیز بازی می‌کند بازمی‌گردد (شکل ۲-۵). در امواج بازگشتی دو کمیت زمان رفت و برگشت امواج ارسالی به دیواره چاه و دامنه امواج بازگشتی به گیرنده اندازه‌گیری می‌شود.



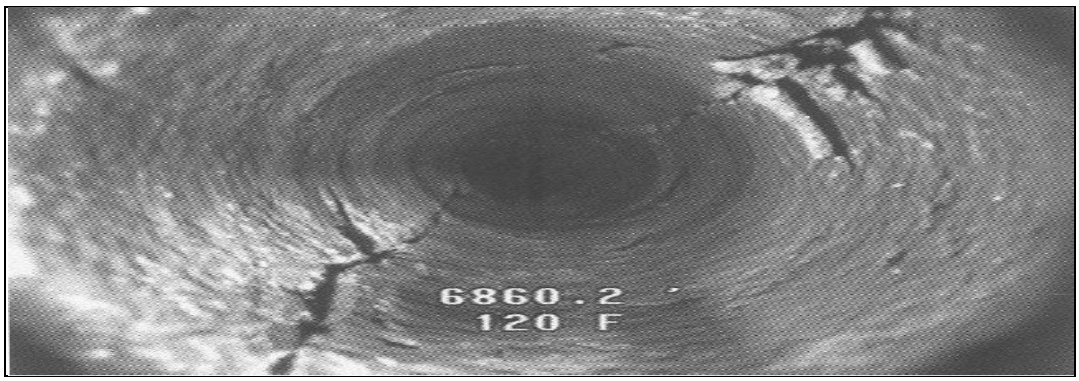
شکل ۲-۵ اصول کلی کار دستگاه‌های تصویربردار صوتی و انعکاس امواج از دیواره چاه [۲۷]

[1] Piezoelectric transducer

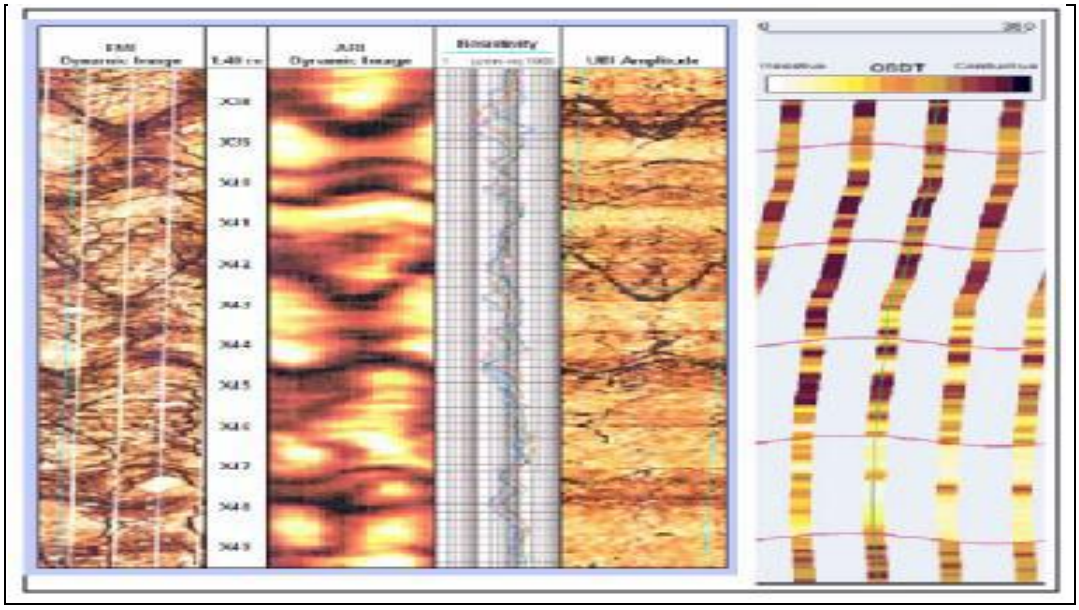
مزیت برجسته‌ای که این دستگاه‌های صوتی نسبت به الکتریکی دارند، قابلیت آن‌ها در تصویربرداری در حفاری‌های گل پایه نفت مانند پایه آب است.

۲-۵-۳- تصویربرداری نوری از چاه

تصویربرداری نوری، قدیمی‌ترین نوع تصویربرداری در چاه‌ها است. دستگاه‌های اولیه آن‌ها دوربین‌های ساده عکاسی بودند که به‌وسیله یک کابل به عمق چاه فرستاده می‌شدند. در سیتم‌های اولیه پس از عکسبرداری تصاویر به سر چاه ارسال می‌شدند تا این‌که سیستم فیلمبرداری وارد صنعت شد [۲۷]. مدل‌های متفاوتی از این دستگاه‌ها با تصاویر رنگی یا سیاه و سفید با کیفیت بالا و قدرت چرخش کامل دوربین عرضه شده است (شکل ۲-۶). این دوربین‌ها در چاه‌های قائم به‌کار گرفته شده‌اند، اما می‌توان از طریق لوله مغزی در چاه‌های افقی نیز به‌کار روند [۲۸]. گل حفاری مورد استفاده در چاه-های باز معمولاً شفاف بوده، درنتیجه امکان استفاده از این ابزارها در این‌گونه چاه‌ها فراهم می‌شود. حتی در تصویربرداری‌های در حین حفاری که سیال حفاری آب است امکان استفاده از این دستگاه‌ها وجود دارد. درجه تفکیک تصاویر حاصل از این دوربین‌ها حداقل به اندازه یک درجه از سایر دستگاه‌ها بالاتر است [۲۷].



شکل ۲-۶ تصویر خاکستری گرفته شده از چاه به‌وسیله دوربین‌های نوری [۲۷]



شکل ۲-۷ مقایسه چاه‌نمودار تصویری الکتریکی و صوتی (چاه‌نمودار تصویری الکتریکی به‌خاطر حساسیت بیشتر سازند اطلاعات ریزتر و بیشتری را ثبت کرده است.) [۲۸]

۲-۶- شکستگی

هر نوع گسیختگی یا جدایش فیزیکی سنگ که ناشی از فزونی تنش‌ها بر مقاومت سنگ باشد که ممکن است به دلیل مکانیزم‌های طبیعی، عوامل حفاری و... در سنگ صورت گیرد شکستگی[۱] نامیده می‌شود.

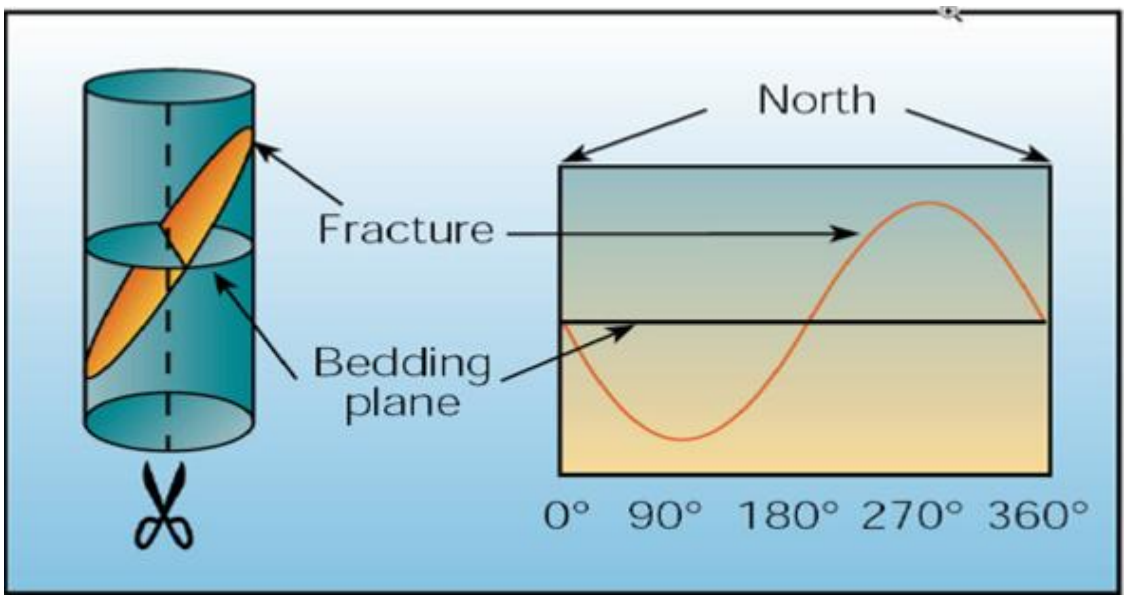
شکستگی‌های طبیعی در مخازن شامل شکستگی‌های بزرگ مقیاس و کوچک مقیاس هستند. شکستگی‌های بزرگ مقیاس مانند گسل‌ها در مقاطع لرزه‌ای مشخص بوده، در حالیکه برای شناسایی شکستگی‌های کوچک مقیاس اطراف چاه، بیشتر روش‌ها مانند نمودارهای چاه‌نگاری، آزمایش چاه، اطلاعات هرزوی گل به دلیل قدرت تفکیک پایین دقت کافی را ندارند. توصیف مغزه‌ها، روش مناسبی برای پی‌بردن به پدیده‌های ریز دیواره چاه است. با این حال مغزه‌ها نیز برای شناسایی شکستگی‌های دیواره چاه دارای محدودیت‌هایی مثل ضریب بازیافت پایین در زون‌های شکسته و هزینه بالاو... هستند [۲۸].

[۱] Fracture

به دلیل اهمیت بالای شکستگی‌های کوچک مقیاس در مخازن به طور طبیعی شکسته شده، تکنولوژی جدیدی برای شناسایی شکستگی‌ها در صنعت نفت به نام نمودارهای تصویری پا به عرصه صنعت نفت گذاشتند [۲۸].

۲-۶-۱- شکستگی درچاه‌نمودارهای تصویری

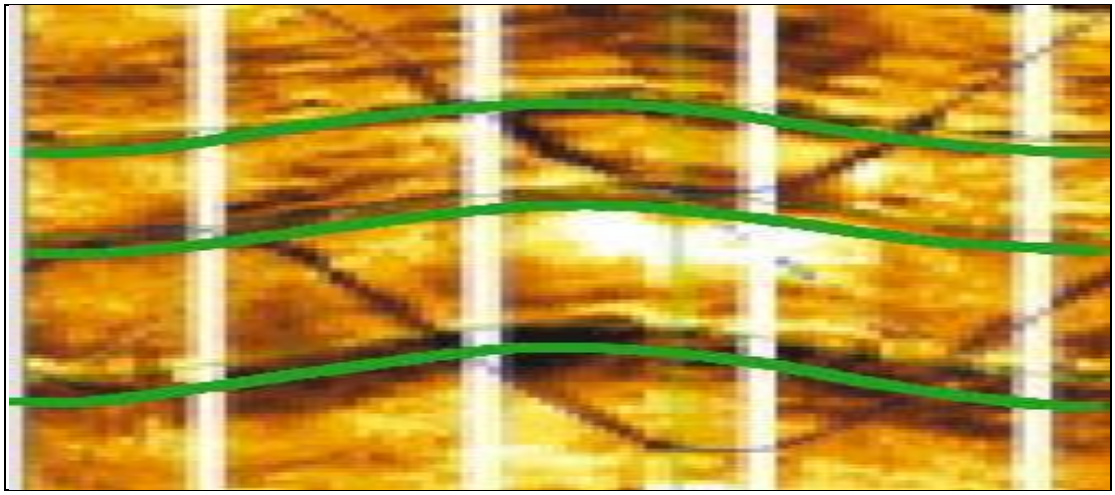
چاه‌نمودارتصویری، تصویری استوانه‌ای از دیواره چاه است. هر پدیده صفحه‌ای شکل، مانند لایه‌بندی و شکستگی که چاه را به‌صورت غیر قائم قطع کرده باشد، در استوانه چاه به شکل بیضی دیده می‌شود. در صورتی که استوانه در امتداد محورش بریده و باز شود یعنی همان شکلی که در نمودار تصویری نمایش داده می‌شود شکستگی یا لایه‌بندی به‌صورت یک موج سینوسی ظاهر می‌شود (شکل۲-۸) [۲۸].



شکل ۲-۸ شکستگی در یک چاه که به‌صورت موج سینوسی دیده می‌شود. [۳۰]

الف) شکستگی‌های باز در چاه‌نمودارتصویری الکتریکی در گل‌های رسانا

دهانه این شکستگی‌ها به‌وسیله گل حفاری پر می‌شود و اگر گل رسانا باشد مقاومتی که در این قسمت توسط نمودار تصویری ثبت می‌شود بسیار کمتر از زمینه سنگ است بنابراین شکستگی باز به‌صورت یک موج سینوسی کامل یا ناپیوسته و تیره رنگ در نمودار تصویری دیده می‌شود (شکل ۲-۹). این شکستگی‌های باز ممکن است در تمام بالشتک‌ها ظاهر نشوند. حداقل عرض قابل مشاهده شکستگی به اندازه قطر یک الکترود است (۵ الی ۶ میلی‌متر) [۲۸].

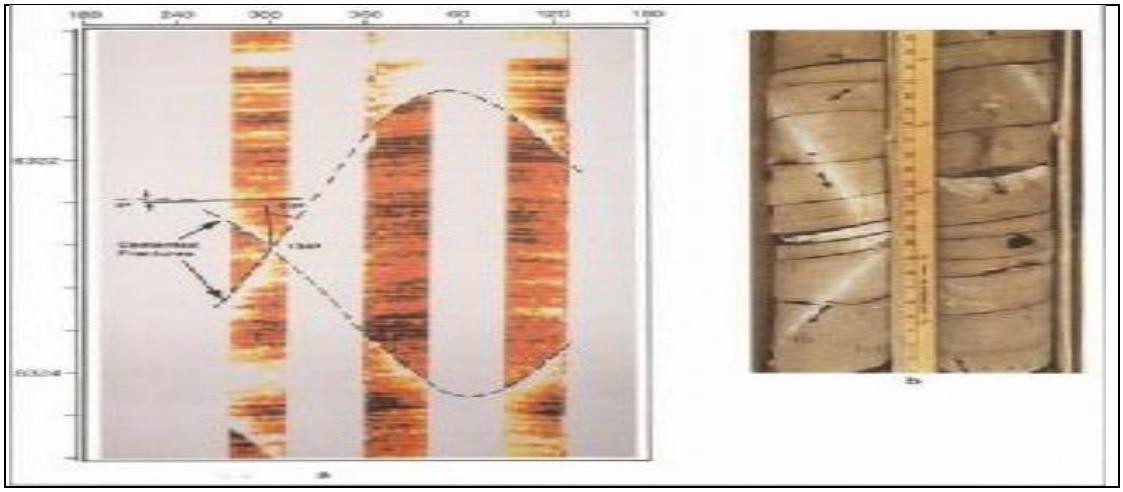


شکل ۲-۹ شکستگی باز در چاه‌نمودار تصویری الکتریکی در گل‌های رسانا به‌صورت موج‌های سینوسی تیره و سینوسی‌های سبز شده لایه‌بندی [۲۸]

ب) شکستگی‌های پرشده در چاه‌نمودارتصویری الکتریکی در گل‌های رسانا

شکستگی‌های پرشده به‌وسیله کانی‌ها، به‌صورت پدیده‌های ریز یا بسیار ریز، کشیده و تقریباً مستقیم با مقاومت الکتریکی بالا (رنگ روشن) و به شکل قائم یا مایل، در چاه‌نمودارهای تصویری دیده می‌شوند (شکل۲-۱۰). ضخامت حداقل شکستگی‌ها که توسط این چاه‌نمودارها قابل شناسایی است به اندازه قطر یک الکترود است (۵ یا ۶ میلی‌متر). علاوه بر آن، زمانی‌که دستگاه از یک شکستگی پرشده عبور می‌کند یک اختلاف مقاومت (تیره یا روشن) بین دو طرف صفحه شکستگی به دلیل پراکندگی

جریان تزریق شده به دیواره چاه در محل شکستگی‌های مایل به‌وجود می‌آید که به آن اثر هاله‌ای گفته می‌شود که یکی از علائم شناسایی شکستگی پرشده است (شکل ۲-۱۰) [۲۸]. ما در این مستند از شکستگی پر شده با عنوان شکستگی بسته نیز نام خواهیم برد.



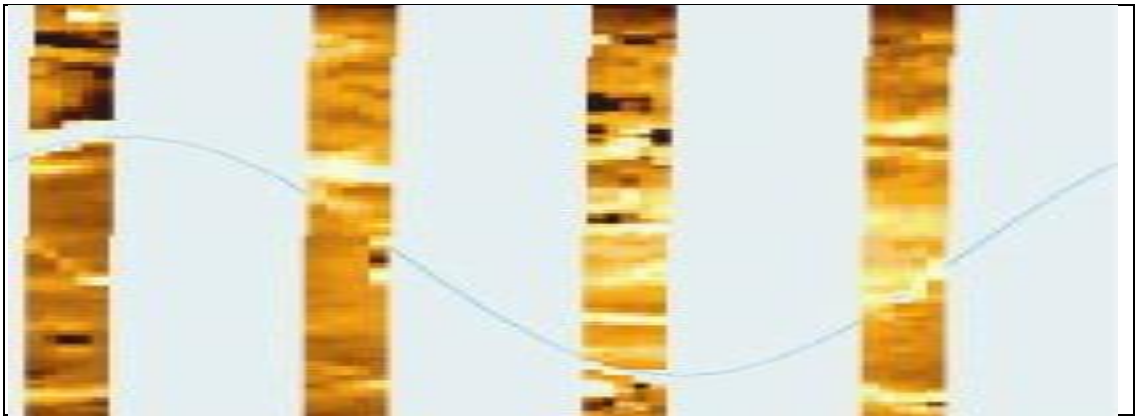
شکل ۲-۱۰ شکستگی پرشده درچاه‌نمودار تصویری الکتریکی در گل‌های رسانا و اثر هاله‌ای که در تصویر مغزه نیز به خوبی دیده می‌شوند [۲۸]

ج) شکستگی‌های باز و پرشده در چاه‌نمودار تصویری الکتریکی در گل‌های نارسانا

دستگاه OBMI[۱] برای تصویربرداری از دیواره چاه در گل‌های نارسانا (نفت پایه) به‌کار می‌رود. در این گل‌ها به دلیل مقاومت بالای گل نمی‌توان از دستگاه‌های معمول تصویربرداری مانند FMI و FMS استفاده کرد [۲۸]. بنابراین از تکنولوژی دیگری برای تزریق جریان به سازند استفاده می‌شود. دستگاه تصویربردار OBMI جریان الکتریکی را با فرکانس بسیار بالا وارد سازند کرده و از طریق محاسبه مقاومت سازند از دیواره چاه تصویربرداری می‌کند.

[۱] Oil base micro imager

شکستگی‌ها در این نمودار متفاوت از FMI و FMS است. در این نمودارها شکستگی‌های باز به رنگ روشن دیده می‌شوند (شکل۲-۱۱). زیرا در شکستگی باز گل حفاری با مقاومت الکتریکی بالا وارد شکستگی شده و در نمودار به‌صورت رنگ روشن ظاهر می‌شود. همچنین شکستگی‌های پرشده نیز به‌صورت روشن دیده می‌شوند زیرا ماده پرکننده شکستگی مقاومت بیشتری از زمینه سازند داشته و در نمودار تصویری روشن‌تر از زمینه دیده می‌شود (شکل ۲-۱۱). در نمودارهای تصویری الکتریکی در گل‌های نارسانا مانند OBMI شکستگی‌های باز و پرشده ظاهر یکسانی داشته و برای تفکیک آن‌ها از هم از نمودار صوتی UBI[۱] استفاده می‌شود [۲۸].



شکل ۲-۱۱ شکستگی باز در چاه‌نمودار تصویری الکتریکی در گل‌های نارسانا [۲۸]

د) شکستگی باز و پرشده در چاه‌نمودار تصویربردار صوتی

در دستگاه تصویربردار UBI سیگنال‌هایی مافوق صوت به سمت دیواره چاه فرستاده می‌شود. زمان رفت و برگشت سیگنال فرستاده شده و دامنه سیگنال برگشتی ثبت می‌شود. دامنه‌های ثبت شده تبدیل به تصویر دیواره چاه و زمان رفت و برگشت تبدیل به شکل دیواره چاه می‌شود. در نمودار UBI شکستگی‌های باز به رنگ تیره دیده می‌شوند. زیرا گل که شکستگی باز را پر می‌کند باعث کاهش

[۱] Ultrasonic borehole imager

دامنه سیگنال می‌شود (شکل۲-۱۲). شکستگی‌های پرشده به رنگ روشن‌تر دیده می‌شوند زیرا ماده پرکننده شکستگی‌ها معمولا متراکم و چگال است و کمتر باعث تضعیف دامنه سیگنال می‌شود [۲۸].

شکل ۲-۱۲ شکستگی باز در چاه‌نمودار تصویری صوتی UBI که به رنگ تیره دیده می‌شود [۲۸]

۲-۷- چاه‌نمودارهای پتروفیزیکی

همانگونه که پیش‌تر در قسمت ۱-۲ نیز بیان شد استفاده از چاه‌نمودارهای تصویری در بسیاری از موارد میسر نمی‌باشد که بعنوان یکی از دلایل آن می‌توان به قطر کم چاه‌هایی اشاره نمود که در گذشته حفر شده‌اند و تعداد آن‌ها نیز در برخی از میادین قابل توجه می‌باشد.

از طرف دیگر اجرای انواع دیگری از لاگ‌برداری که چاه‌نمودارهای پتروفیزیکی نامیده می‌شود در این چاه‌ها امکان‌پذیر می‌باشد که هدف این تحقیق نیز در نهایت استفاده از این لاگ‌ها به منظور شناسایی شکستگی‌ها می‌باشد. لذا در این قسمت لازم است به معرفی این لاگ‌ها پرداخته شود.

پتروفیزیک[۱] علمی است که به بررسی خواص فیزیکی مجموعه ی سنگ و سیال در مقابل تحریک های فیزیکی، امواج الکترومغناطیسی، الکتریکی، پرتوهای رادیواکتیو، امواج مکانیکی و ... می‌پردازد.

[۱] Petrophysics

داده های به دست آمده از عملیات پتروفیزیک که شامل ثبت پاسخ به تحریکهای فیزیکی فوق است داده‌های پتروفیزیکی نامیده می‌شوند.

در مستند پیش رو علاوه بر از لاگ و یا داده‌ی پتروفیزیکی از عبارت چاه‌نمودار[۱] نیز استفاده گردیده است. چاه‌نمودارهای بکار گرفته شده در این تحقیق به شرح زیر می‌باشند.

CALI: این چاه‌نمودار که نام کامل آن Caliper می‌باشد، یک اندازه‌گیری پیوسته از اندازه و شکل حفره‌ی چاه را ارائه می‌دهد.

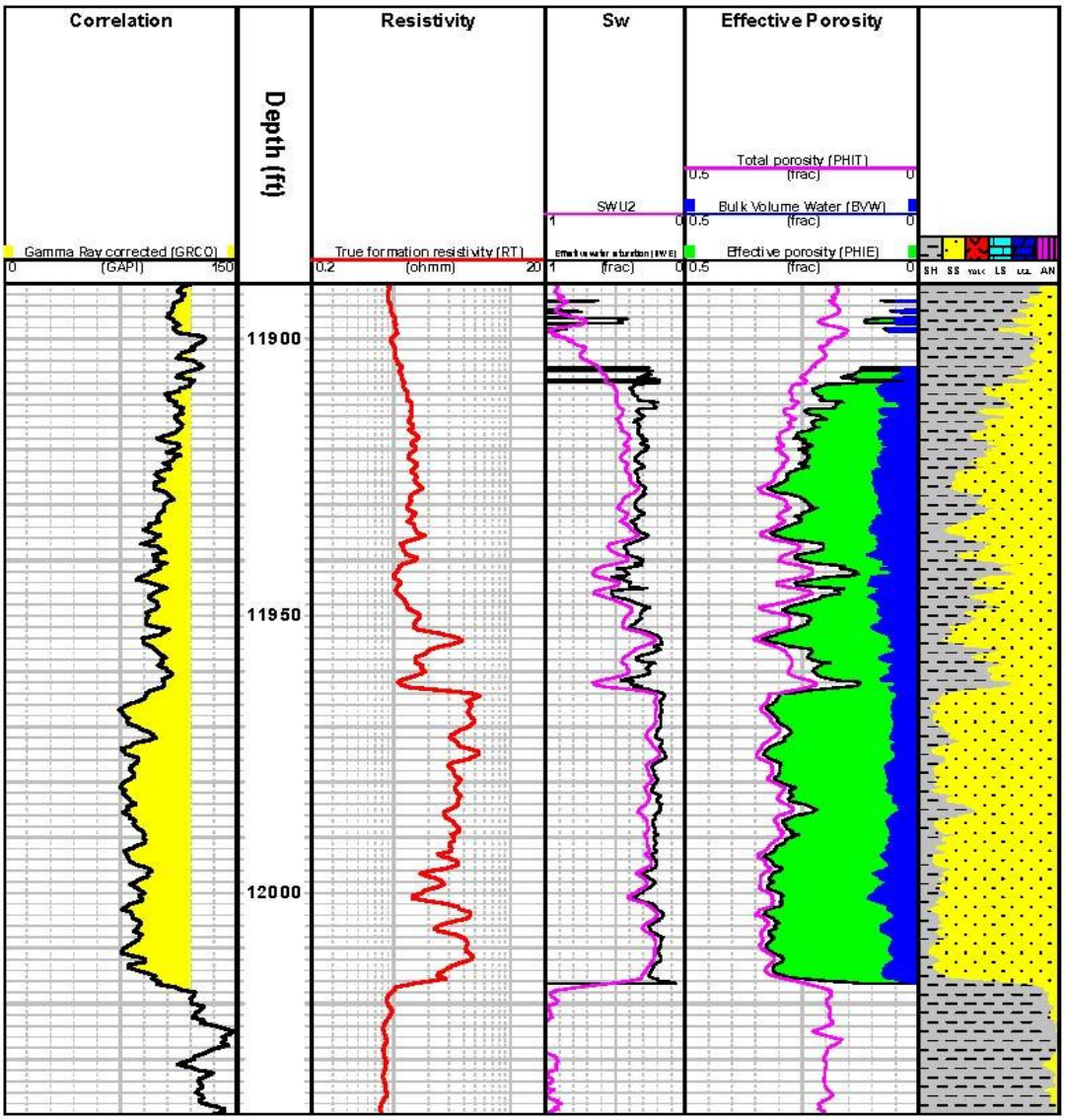
CGR: این واژه مخفف عبارت Computed Gamma Ray بوده و عبارت است از مجموع میزان بازتاب‌های پتاسیم[۲] و توریم[۳].

DT: چاه‌نمودار سونیک[۴] در پتروفیزیک از اهمیت فوق‌العاده‌ای برخوردار است. از طریق چاه‌نمودار سونیک می‌توان به محاسبه‌ی میزان تخلخل[۵] سنگ پرداخت. ایده‌ی این ابزار بسیار ساده می‌باشد و آن اینست که ما از طریق امواج صوتی به میزان تخلخل پی می‌بریم چرا که هر چه تخلخل افزایش یابد سرعت نفوذ موج بیشتر می‌گردد.

NPHI: این چاه‌نمودار مورد استفاده در صنعت نفت و گاز می‌باشد و نام کامل آن Neutron Porosity. این چاه‌نمودار نیز بسیار مورد استفاده در صنعت بوده و داده‌ی آن توسط ابزاری ایجاد می‌گردد که از تکنیک لاگ برداری نوترونی استفاده می‌کند. این لاگ نیز در تعیین میزان تخلخل سازه مورد استفاده دارد.

---

[۱] Well Log
[۲] Potassium
[۳] Thorium
[۴] Sonic
[۵] Porosity



شکل ۲-۱۳ نمونه‌ای از داده‌های پتروفیزیکی

PEF: نام کامل این لاگ Photo Electric Factor می‌باشد. ابزار تهیه‌ی این لاگ شامل یک رشته سیم رسانا می‌باشد.

RHOB: نام رسمی این لاگ Bulk Density می‌باشد. این لاگ به طور رایج در صنعت مورد استفاده قرار می‌گیرد و معمولا نمایانگر چگالی سنگ می‌باشد.

SGR: این لاگ میزان طبیعی تشعشع گاما را اندازه‌گیری می‌کند. تحلیل منبع تشعشع طبیعی گاما به کارشناسان کمک می‌کند اطلاعاتی از اجزای تشکیل دهنده‌ی سازه بدست آورند.

SWE: میزان جذب آب موٴثر در عمق‌های مختلف توسط این لاگ تصویر می‌شود.

CALCITE: این لاگ میزان کالسایت[۱] موجود در هر عمق را اندازه‌گیری می‌کند. کالسایت در اغلب سازه‌ها یافت می‌شود.

SHALE: این لاگ وظیفه‌ی اندازه‌گیری میزان شیل[۲] موجود در هر عمق چاه را بعهده دارد.

[۱] Calcite
[۲] Shale

فصل سوم

مفاهیم داده‌کاوی

۳-۱- مقدمه

در طول اجرای این پروژه برای بدست آوردن نتایج مطلوب تعداد زیادی از روش‌های پیش‌پردازش و روش‌های دسته‌بندی داده‌ها استفاده شده است که در نهایت از میان آنها تنها تعداد محدودی از الگوریتم‌ها در روش پیشنهادی ما مورد استفاده قرار خواهند گرفت. در این فصل بیشتر سعی شده به توضیح الگوریتم‌هایی پرداخته شود که در نهایت بعنوان بخشی از روش پیشنهادی ما انتخاب شده‌اند.

۳-۲- پیش پردازش داده‌ها

در کلیه‌ی مسائل داده‌کاوی برای اینکه الگوریتم‌های دسته‌بندی به صورت کارامدتر عمل کنند نیاز به این داریم که داده‌های اولیه‌ی مسأله را مورد پردازش و آماده‌سازی‌های اولیه قرار دهیم. که ما در ادامه‌ی این بخش به معرفی چند نمونه‌ی خاص از روش‌های پیش‌پردازش می‌پردازیم.

۳-۲-۱- پنجره‌گذاری

تکنیک پنجره‌گذاری[۱] را می‌توان به دو منظور عمده به کار گرفت یکی کاهش حجم داده‌های مسأله و در نتیجه بهبود عملکرد مراحل بعدی داده‌کاوی که ما در این پروژه به این کار نیاز نداریم و دیگری آماده سازی و ساده‌سازی داده‌ها برای آنکه ماشین‌های دسته‌بندی عملکرد بهتری را داشته باشند [۳۱]. پنجره‌گذاری یک تکنیک رایج در برخی از زمینه‌های داده‌کاوی از جمله در استخراج ویژگی در پردازش سیگنال‌های صوتی و گفتار [۳۲]، سیگنال‌های بیولوژیکی [۳۳]و [۳۴] و... کاربرد دارد.

[۱] Windowing

پنجره‌گذاری از نظر چگونگی انتخاب طول پنجره دو نوع مختلف دارد که در اینجا تنها لازم است به نوع ساده‌تر آن یعنی پنجره‌ی با طول ثابت[۱] بپردازیم. برای استفاده از این تکنیک لازم است داده‌های هر مسأله‌ی داده‌کاوی را در کنار هم و به عنوان یک سیگنال در نظر گرفت که در این صورت می‌توان یک پنجره‌ی با طول ثابت را بر روی آن لغزانید، در هر گام از این لغزش می‌توانیم یک ویژگی سیگنال درون پنجره را محاسبه نموده و مقدار بدست آمده را جایگزین یکی از داده‌های درون پنجره بنمائیم.

دو سؤال در اینجا مطرح می‌شود یکی اینکه کدام ویژگی سیگنال درون پنجره باید محاسبه شود، که در پاسخ باید گفت، بسته به داده‌های مسأله ممکن است هریک از ویژگی‌های سیگنال از قبیل گشتاورهای ریاضی، ضرائب فوریه، ضرائب موجک[۲] و سایر توابع استفاده نمود و ما در قسمت بعد به توضیح یکی از این ویژگی‌ها یعنی گشتاورهای ریاضی می‌پردازیم.

دومین سؤالی که در اینجا مطرح می‌شود اینست که کدامیک از داده‌های درون پنجره باید با مقدار بدست آمده بعنوان ویژگی سیگنال پنجره جایگزین شود، که در پاسخ باید گفت که باز هم با توجه به داده‌های مسأله این مورد می‌تواند متغیر باشد ولی به صورت عمده سه نوع جایگزینی مطرح می‌شود. جایگزینی با داده‌ی اول پنجره که به آن نگاه به آینده نیز اطلاق می‌شود. جایگزینی با داده‌ی وسط پنجره، و در آخر جایگزینی با داده‌ی پایانی درون پنجره.

۳-۲-۲- گشتاورهای ریاضی

هر توضیع آماری قابل توصیف با تعدادی مشخصه از قبیل میانگین، واریانس و کج‌شدگی[۳] می‌باشد، و ممان‌های یک توضیع از یک متغیر تصادفی نیز با این مشخصه‌ها در ارتباط هستند.

---

[۱] Fixed Sliding Window

[۲] Wavelet

[۳] Skewness

اولین ممان یا بعبارتی اولین ممان حول صفر بعنوان میانگین[۱] شناخته شده می‌باشد. در درجه‌های بالاتر ممان اغلب به صورت ممان حول میانگین مورد استفاده‌ی بیشتری دارند چرا که اطلاعات واضح-تری را درباره‌ی شکل توضیع در اختیار قرار می‌دهند [۳۵].

nامین ممان یک تابع پیوسته مقدار حقیقی f(x) حول c به صورت زیر مطرح می‌شود:

$$\mu'_n = \int_{-\infty}^{\infty} (x - c)^n f(x) dx \qquad (۱-۳)$$

۳-۲-۲-۱- واریانس

دومین گشتاور مرکزی همان واریانس می‌باشد که ریشه دوم مثبت آن انحراف از معیار نامیده می‌شود و با σ نشان داده می‌شود.

۳-۲-۲-۲- گشتاورهای ریاضی نرمال

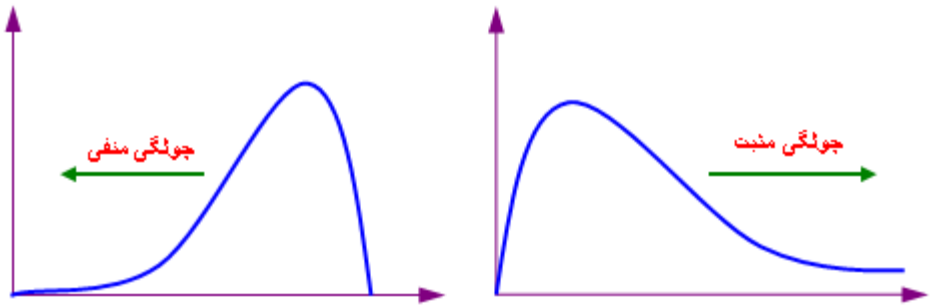
nامین گشتاور مرکزی نرمال شده و یا ممان استاندارد شده[۲] به nامین گشتاور مرکزی اطلاق می‌شود که بر σ^n تقسیم شده باشد، $x = E((x-\mu)^n)/\sigma$. این گشتاورهای مرکزی نرمال شده مقادیر بدون بعد می‌باشند که بیان‌کننده‌ی یک توضیع بدون تاثیرپذیری از هر تغییر خطی در مقیاس می-باشند [۳۵].

۳-۲-۲-۳- چولگی[۳]

سومین گشتاور مرکزی بیان کننده اندازه‌ی ناهماهنگی توضیع می‌باشد، هر توضیع متقارنی که دارای گشتاور سوم باشد چولگی برابر با صفر خواهد داشت. گشتاور سوم نرمال شده چولگی نامیده می‌شود

[۱] Mean
[۲] Standardised Moment
[۳] Skewness

که معمولا با ۷ نشان داده می‌شود. یک توضیع اگر دارای چولگی به سمت چپ باشد (ادامه‌ی منحنی توضیع در سمت چپ بیشتر باشد) آنگاه دارای چولگی منفی خواهد بود و برعکس اگر چولگی به سمت راست باشد چولگی توضیع یک مقدار مثبت خواهد داشت.



شکل ۳-۱ چولگی مثبت و چولگی منفی

۳-۲-۲-۴- کشیدگی[۱]

چهارمین گشتاور مرکزی بیان کننده‌ی میزان بلندی و باریکی و یا کوتاهی و پهنی توضیع در مقایسه با توضیع نرمال با همان واریانس می‌باشد. همانطور که از توان چهارم انتظار می‌رود این گشتاور همواره دارای مقدار مثبت می‌باشد. گشتاور چهارم مرکزی در یک توزیع نرمال معادل $۳\sigma^{۴}$ می‌باشد. کشیدگی k به صورت گشتاور مرکزی چهارم منهای ۳ تعریف می‌گردد. اگر یک توضیع دارا یک قله‌ی تیز در میانگین بوده و نیز دارای دنباله‌های ضخیم باشد آنگاه گشتاور چهارم دارای اندازه بزرگ خواهد بود که به آن leptokurtic اطلاق می‌شود. همچنین توضیعی که دارای قله‌ای پهن و دنباله‌هاای باریک دارد دارای اندازه کوچک کشیدگی خواهد بود که platykurtic نامیده می‌شود [۳۶].

$$\beta_{۲} = \frac{E[(x-\mu)^{۴}]}{(E[(X-\mu)^{۲}])^{۲}} = \frac{\mu_{۴}}{\sigma_{۴}} \qquad (۲-۳)$$

[۱] Kurtosis

۳-۲-۳- نرمال‌سازی داده‌ها

برای آنکه تاثیرگذاری داده‌ها بر روی نتایج کلاس‌بندها به صورت غیرواقعی افزایش و یا کاهش نیابد باید به عنوان یک گام پیش‌پردازش و قبل از اعمال الگوریتم‌های دسته‌بندی بر روی داده‌ها، داده‌ها را نرمال‌سازی کنیم. نکته‌ای که در این گونه مسائل حائز اهمیت است این است که داده‌های آموزش و داده‌های آزمون باید تواما نرمالسازی شوند و نه به صورت جداگانه.

تا کنون چندین روش نرمال‌سازی در مباحث مربوط به علم آمار مطرح شده است که ما در اینجا به یک مورد از پرکاربردترین آنها در داده‌کاوی اشاره می‌کنیم.

در این روش نرمال‌سازی کلیه‌ی داده‌ها به بازه‌ی میان صفر و یک نگاشت می‌شوند که بنابراین به آن unity-based normalization نیز اطلاق می‌گردد و نحوه‌ی محاسبه‌ی آن به شکل زیر می‌باشد: [۳۷]

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (۳-۳)$$

۳-۲-۴- Info Gain Attribute Evaluation

این الگوریتم برای سنجش میزان ارزش یک مشخصه[۱] بر اساس بهره‌ی اطلاعاتی[۲] که این ویژگی در برابر کلاس مورد نظر دارد، مورد استفاده قرار می‌گیرد [۳۸]. ما در قسمتهایی از روند اجرای پروژه بر آن شدیم تا نمودارهای مختلف (که هر یک به منزله‌ی یک ویژگی محسوب می‌شوند) را بر اساس میزان اهمیت و تاثیرگذاری بر روی الگوریتم کلاس‌بند نمره دهی نمائیم، که در این مهم از الگوریتم InfoGainAttributeEval استفاده نمودیم. در اینجا لازم به ذکر است که این الگوریتم در نرم‌افزارهای داده‌کاوی Weka و Rapid Miner نیز پیاده سازی شده است.

[۱] Attribute

[۲] Information Gain

نحوه کار این این الگوریتم چیزی به جز استفاده از مفهوم بهره اطلاعات و آنتروپی[1] نمی‌باشد که در ادامه مروری بر ماهیت و نحوه محاسبه این دو مفهوم خواهیم داشت.

۳-۳- تئوری اطلاعات

تئوری اطلاعات در بردارنده کاربردهایی می‌باشد که از آن‌جمله می‌توان به اندازه‌گیری اطلاعات اشاره نمود. اندازه‌گیری اطلاعات نخستین بار توسط کلاد شانون[2] مطرح گردید [۳۹].

۳-۳-۱- آنتروپی

در تئوری اطلاعات، آنتروپی عبارت است از یک نوع اندازه‌گیری میزان تردید[3] در یک متغیر تصادفی. آنتروپی در کاربردهای مختلف می‌تواند به گونه‌های متفاوت بیان شود، برای مثال فرض کنید می-خواهیم با یک درحت تصمیم یک مسأله‌ی دسته‌بندی را حل نمائیم [۴۰].

در مثال مورد نظر، درختی را به صورت بالا به پائین ایجاد خواهیم نمود که در هر گره‌ی آن یکی از مشخصه‌های داده‌ها مورد سنجش قرار می‌گیرد. اما سؤال اساسی که در اینجا مطرح است این است که کدام مشخصه را برای تجزیه در هر گره انتخاب کنیم؟ پاسخ اینست که باید مشخصه‌ای انتخاب شود که بتواند مسیر جستجو را کوتاه‌تر نموده و سریعتر ما را به گره‌های خالص هدایت نماید (منظور از گره خالص در واقع همان برگ است که نماینده‌ی تنها یک کلاس خاص می‌باشد، در حالی که گره ناخالص ممکن است مسیر جستجو را به سمت کلاس‌های مختلفی هدایت نماید.).

---

[1] Entropy
[2] Claud Shannon
[3] Uncertainty

اندازه این خالص بودن را در اینجا اطلاعات[1] می‌نامیم. بنابراین اطلاعات عبارت است از میزان اطلاعاتی که برای تشخیص کلاس یک نمونه جدید لازم داریم و محاسبه آن بر اساس تعداد کلاس‌های آقا و خانمی هست که در ذیل هر گره قرار می‌گیرد.

عکس آنچه که در بالا مطرح شد مفهوم آنتروپی عبارتست از میزان ناخالصی گره‌ها. با فرض آنکه دو کلاس a و b داریم فرمول محاسبه آنتروپی به شکل زیر خواهد بود:

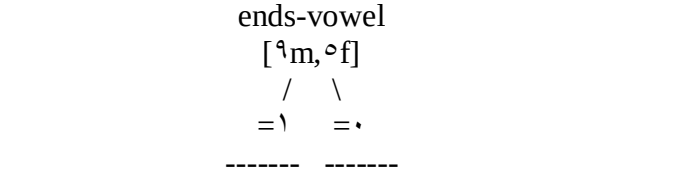
Entropy = - p(a)*log(p(a)) - p(b)*log(p(b))

(۴-۳)

که در آن p به معنای احتمال می‌باشد. این فرمول برگرفته از فرمول شانون می‌باشد که پرداختن به آن و نحوه‌ی اثباتش از حوصله‌ی این مستند خارج می‌باشد.

۳-۳-۲- بهره‌ی اطلاعات

فرض کنید می‌خواهیم با استفاده از یک درخت تصمیم تشخیص دهیم که یک نام متعلق به دسته‌ی آقایان است یا دسته‌ی خانم‌ها، یکی از مشخصه‌ها می‌تواند صدا دار بودن حرف پایانی باشد، حالا فرض کنید در مجموع ۱۴ نام(نمونه) داشته باشیم که ۵ تای آنها متعلق به خانم و ۹ تای آن متعلق به آقا باشد. ولی وقتی صفت مورد نظر را اعمال میکنیم ممکن است به یکی از دو حالت جدید برویم، حالت اول اینکه ۳آقا و ۴خانم داریم و در صورتی رخ می‌دهد که حرف پایانی صدادار باشد، حالت دوم حالتیست که در آن ۶ نام آقا و ۱ خانم وجود دارد و زمانی رخ می‌دهد که حرف پایانی صدادار نباشد. یک قطعه از درخت که این گفته‌ها را تصویر می‌کند می‌تواند به شکل زیر باشد:



[1] Information

[۳m,٤f] [٦m,۱f]

پس قبل از تجزیه توسط صفت مورد نظر داشتیم: p(m)=۹/۱٤ , p(f)=٥/۱٤

و بنابر تعریف آنتروپی، برای محاسبه آنتروپی قبل از تجزیه داریم:

Entropy_before = -٥/۱٤*log(٥/۱٤) - ۹/۱٤*log(۹/۱٤) = ...

به همان صورت بعد از تجزیه نیز به ترتیب برای برای آنتروپی سمت چپ و سمت راست درخت داریم:

Entropy_left = -۳/۷*log(۳/۷) -٤/۷*log(٤/۷) = ...
Entropy__right = -٦/۷*log(٦/۷) -۱/۷*log(۱/۷) = ...

حال باید آنتروپی چپ و راست را با هم ترکیب کنیم که برای این کار باید وزن هر یک را نیز در محاسبات وارد نمائیم که هر دو برابر و معادل با ۷/۱۴ هستند، چرا که تعداد ۷ نمونه به هر یک از دو گره چپ و راست راه پیدا کرده‌اند.

Entropy_after = ۷/۱٤*Entropy_left + ۷/۱٤*Entropy__right = ...

حال با مقایسه آنتروپی قبل و پس از تجزیه می‌توانیم بهره اطلاعات یا به عبارتی اندازه اطلاعاتی که از انجام تجزیه بر اساس یک صفت خاص به دست می‌آید را محاسبه کنیم:

Information_Gain = Entropy_before - Entropy_after (۵-۳)

۳-۴- دسته‌بندی داده‌ها

۳-۴-۱- چگونگی انتخاب الگوریتم دسته‌بندی

در هر مسئله‌ی یادگیری ماشین برای آنکه یک الگوریتم دسته‌بندی ویا بعبارتی کلاسیفایر مناسب را انتخاب نمائیم اولین پرسشی که باید به دنبال پاسخ آن باشیم این است که به چه میزان داده در دست است؟ هیچ؟ خیلی کم؟ زیاد؟ خیلی زیاد که هر روز هم به حجم آن افزوده می‌شود؟ ما در طول روال بررسی کلاسیفایرهای مختلف سعی داشتیم به یک جمع‌بندی در مورد اصول انتخاب کلاسیفایر مناسب برسیم که عبارتند از:

۱. همه‌ی کلاسیفایرها در مقابل عدم توازن کلاس‌ها دچار کاهش عملکرد می‌شوند. این بدان معناست که داده‌های این تحقیق کلاسیفایر را دچار کاهش عملکرد خواهد نمود. چرا که این داده‌ها به شکلی تقسیم بندی شده‌اند که بیش از ۸۵ درصد از داده‌ها در یک کلاس و کمتر از ۱۵ درصد داده‌ها در دو کلاس دیگر قرار می‌گیرند.
۲. شبکه‌های عصبی برای آموزش به حجم داده‌ی زیاد نیاز دارند بعلاوه‌ی اینکه نسبت به عدم توازن کلاس‌ها ضعف بیشتری را از خود نشان می‌دهند. پس برای این تحقیق نمی‌توانند انتخاب مناسبی باشند که البته آزمایشات ما نیز این موضوع را تائید نمود.
۳. الگوریتم‌های استنتاج درخت تصمیم حتی با داده‌های آموزشی کم می‌توانند عملکرد خوبی داشته باشند ولی با این شرط که داده‌های آموزشی بتوانند نماینده‌ی خوبی از جامعه‌ی آماری خود باشند.
۴. اگر حجم زیاد و یا نامحدودی از داده‌ها(داده‌های برچسب دار) در دسترس باشد چندان تفاوتی ندارد که از چه کلاسیفایری استفاده نمائیم بلکه در این مواقع بیشتر باید به مواردی از قبیل سرعت آموزش، سرعت اجرا، قابلیت استخراج فرمول کلاسبندی و مواردی از این دست

توجه نمود که وابسته به صورت مسأله‌ی مورد مطالعه می‌باشند. البته داده‌های مسأله‌ی مورد نظر پروژه در این دسته قرار نمی‌گیرند چرا که ما تعداد بسیار محدودی از داده‌ها را در اختیار داریم.

۵. در مواردی که تعداد نمونه‌ها (نمونه‌های دارای برچسب) محدود و کم هستند بنا بر تئوری یادگیری ماشین بهتر است از کلاسیفایرهای با بایاس[۱] بالا و با واریانس[۲] کم استفاده شود و برعکس هر چه حجم داده‌های نمونه بیشتر باشد بهتر است از کلاسیفایرهای با بایاس کمتر و واریانس بیشتر استفاده شود. برای مثالی از کلاسیفایرهای با بایاس بالا می‌توان به Naive Bayes اشاره نمود که در آزمایشات مربوط به این پروژه نیز پاسخ نسبتا خوبی را از خود نشان داد [۴۱].

۶. نکته‌ی پایانی اینست که پیچیدگی و نیز کارائی راه‌حل یک مسأله‌ی یادگیری ماشین نمی-تواند بعنوان خصوصیت الگوریتم کلاسیفایر مطرح باشد، چرا که علاوه بر کلاسیفایر بخش‌های زیاد دیگری نیز در یک مسأله‌ی یادگیری ماشین لازم به اجرا هستند که از آن میان می‌توان به انتخاب ویژگی، رگلاسیون و ایجاد محدودیت اشاره نمود.

لازم به ذکر است که کارائی بالای الگوریتم Naïve Bayes و Random Tree در کلاس‌بندی مسائلی که دارای حجم داده نسبتا کم هستند از لحاظ تئوری و نیز با آزمایشات عملی به اثبات رسیده است که می‌توان برای نمونه به [۴۲] و [۴۳] اشاره نمود.

این اصول در طول تحقیقات این پروژه از میان منابع مختلف گردآوری شده است که مطمئنا کامل نبوده و خوانندگان محترم می‌توانند در صورت نیاز در این مورد تحقیقات بیشتری را انجام دهند.

---

[۱] Bias

[۲] Variance

۳-۴-۲- الگوریتم دسته‌بندی Naïve Bayes

الگوریتم‌های مبتنی بر روش Naïve Bayes راه حلی مؤثر برای مسائل دسته‌بندی داده هستند که برای داده‌های مفقود شده از مجموعه‌ی داده‌های آموزشی فرض‌هائی را طراحی می‌کند [۴۴]. برای درک نحوه‌ی عملکرد این الگوریتم در ابتدا لازم است با الگوریتم دسته‌بندی Bayes آشنا باشیم، چرا که الگوریتم Naïve Bayes حالت خاصی از آن می‌باشد لذا ابتدا بصورت مختصر به توضیح الگوریتم Bayes می‌پردازیم.

استدلال بیزی روشی بر پایه‌ی احتمالات برای استنتاج کردن می‌باشد و اساس این روش بر این اصل استوار است که برای هر کمیتی یک توزیع احتمال وجود دارد که با مشاهده‌ی یک داده‌ی جدید و استدلال در مورد توزیع احتمال آن می‌توان تصمیمات بهینه‌ای اتخاذ نمود. از نظر اهمیت درک الگوریتم بیزی می‌توان گفت یادگیری بیزی در برخی از کاربردها نظیر دسته‌بندی متن توانسته است راه‌حل‌های عملی مفیدی را ارائه نماید و حتی فراتر از آن، مطالعه‌ی یادگیری بیزی به فهم سایر روش‌های یادگیری که بطور مستقیم از احتمالات استفاده نمی‌کنند نیز بسیار کمک می‌کند.

نگرش بیزی به یادگیری ماشین (و یا هر فرآیند دیگر) به شکل زیر می‌باشد:

۱. دانش موجود درباره موضوع را بصورت احتمالاتی فرموله می‌کنیم.
   a. برای این کار مقادیر کیفی دانش را به صورت توضیع احتمال، فرضیات استقلال و غیره مدل می‌نمائیم که این مدل دارای پارامترهای ناشناخته‌ای خواهد بود.
   b. برای هر یک از مقادیر ناشناخته توضیع احتمال اولیه‌ای در نظر گرفته می‌شود که بازگوکننده‌ی باور ما به محتمل بودن هر یک از این مقادیر بدون دیدن داده است.
۲. داده را جمع‌آوری می‌کنیم.
۳. با مشاهده‌ی داده‌ها مقدار توزیع احتمال ثانویه را محاسبه می‌کنیم.

۴. با استفاده از این احتمال ثانویه:

a. به یک نتیجه‌گیری در مورد عدم قطعیت می‌رسیم.

b. با میانگین‌گیری روی مقادیر احتمال ثانویه پیش‌بینی انجام می‌دهیم.

c. برای کاهش خطای ثانویه مورد انتظار تصمیم می‌گیریم.

از ویژگی‌های این روش که بهتر است در اینجا به آنها اشاره کنیم اینست که:

- مشاهده‌ی هر نمونه می‌تواند بصورت جزئی باعث افزایش و یا کاهش احتمال درست بودن یک فرضیه گردد.
- برای بدست آوردن احتمال یک فرضیه می‌توان دانش قبلی را با نمونه‌ی مشاده شده ترکیب نمود. این دانش قبلی به دو طریق بدست می‌آید: یکی اینکه احتمال قبلی برای هر فرضیه موجود باشد و دیگری اینکه برای داده‌ی مشاهده شده، توزیع احتمال هر فرضیه‌ی ممکن موجود باشد.
- روش‌های بیزی فرضیه‌هایی ارائه می‌دهند که قادر به پیش‌بینی احتمالی هستند(مثل: در این عمق از چاه به احتمال ۸۰ درصد شکستگی وجود ندارد)
- نمونه‌ی جدید را می‌توان با ترکیب وزنی چندین فرضیه دسته‌بندی نمود.

۳-۴-۲-۱- تئوری Bayes

در یادگیری ماشین معمولا در فضای فرضیه‌ی H بدنبال بهترین فرضیه‌ای هستیم که در مورد داده-های آموزشی D صدق کند. یک راه تعیین بهترین فرضیه اینست که بدنبال محتمل‌ترین فرضیه‌ای باشیم که با داشتن داده‌های آموزشی D و احتمال قبلی در مورد فرضیه‌های مختلف می‌توان انتظار

داشت. تئوری بیز چنین راه‌حلی را ارائه می‌دهد که راه‌حل مستقیمی بوده و نیاز به جستجو ندارد [۴۵].

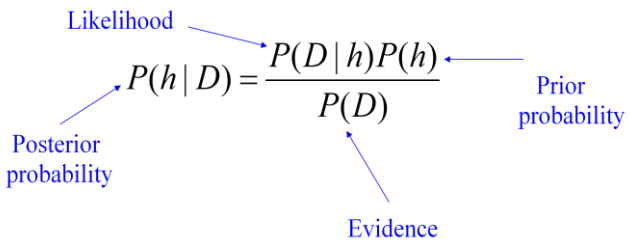
فرض کنید که فضای فرضیه H و مجموعه مثالهای آموزش D موجود باشند .مقادیر احتمال زیر را تعریف میکنیم:

۱. P(h)=احتمال اولیه ای که فرضیه h قبل از مشاهده مثال آموزشی Dداشته است (prior probablity)اگر چنین احتمالی موجود نباشد میتوان به تمامی فرضیه ها احتمال یکسانی نسبت داد.

۲. P(D)=احتمال اولیه ای که داده آموزشی D مشاهده خواهد شد.

۳. P(D|h)=احتمال مشاهده‌ی داده آموزشی D به فرض آنکه فرضیه h صادق باشد.

در یادگیری ماشین علاقه مند به دانستن P(h|D) یعنی احتمال اینکه با مشاهده داده آموزشیD فرضیهh صادق باشد، هستیم. این رابطه احتمال ثانویه (posterior probablity) نامیده می‌شود. توجه شود که احتمال اولیه مستقل از داده آموزشی است ولی احتمال ثانویه تاثیر داده آموزشی را منعکس میکند. سنگ بنای یادگیری بیزی را تئوری بیز تشکیل میدهد .این تئوری امکان محاسبه احتمال ثانویه را بر مبنای احتمالات اولیه میدهد:



شکل ۳-۲ تئوری بیز

همانطور که مشاهده میشود با افزایش P(D) مقدار P(h|D) کاهش می یابد .زیرا هر چه احتمال مشاهدهD مستقل از h بیشتر باشد به این معنا خواهد بود کهD شواهد کمتری در حمایت ازh در بر دارد.

۳-۴-۲-۲- فرضیه با حداکثر احتمال (MAP)

در مسایلی که مجموعه ای از فرضیه هایH وجود داشته و بخواهیم محتمل‌ترین فرضیه را از میان آنان انتخاب بکنیم، فرضیه با حداکثر احتمال Maximum A Posteriori (MAP) hypothesis نامیده میشود و از رابطه ۳-۶ بدست می آید. توجه نمائید که در این رابطه مقدار P(D)مستقل از h بوده و حذف میشود

$$\begin{aligned} h_{MAP} &\equiv \arg\max_{h \in H} P(h \mid D) \\ &= \arg\max_{h \in H} \frac{P(D \mid h) P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D \mid h) P(h) \end{aligned} \qquad (۶-۳)$$

۳-۴-۲-۳- فرضیه‌ی Maximum Likelihood (ML)

در مواقعی که هیچ اطلاعی در مورد P(h) وجود نداشته باشد می‌توان فرض کرد که تمام فرضیه‌های H دارای احتمال اولیه یکسانی هستند .در اینصورت برای محاسبه فرضیه با حداکثر احتمال میتوان فقط مقدار$P(D \mid h)$ را در نظر گرفت .این مقدار liklihood داده D با فرض h نامیده میشود و هر فرضیه ای که مقدار آنرا ماکزیمم کند فرضیهmaximum liklihood (ML) نامیده میشود:

$$h_{ML}=\text{argmax}_{h\in H} P(D \mid h) \qquad (۷-۳)$$

۳-۴-۲-۴- دسته بندی کننده بیزی بهینه[۱]

در عمل محتمل‌ترین دسته بندی برای یک نمونه جدید از ترکیب وزنی پیش بینی تمامی فرضیه‌ها بدست میاید .اگر دسته بندی مثال جدید بتواند هر مقدار v_j از مجموعه V را داشته باشد در اینصورت احتمال اینکه مثال جدید دسته بندی V_j را داشته باشد برابر است با:

$$P(v_j \mid D) = \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) \qquad (۸-۳)$$

مقدار ماکزیمم رابطه فوق دسته بندی بهینه این نمونه را مشخص خواهد نمود:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j \mid h_i) P(h_i \mid D) \qquad (۹-۳)$$

۳-۴-۲-۵- دسته‌بندی کننده‌ی Naïve Bayes

این روش یکی از روش‌های بسیار عملی در یادگیری ماشین می‌باشد این روش در مسائلی کاربرد دارد که:

- نمونه x توسط ترکیب عطفی ویژگیها قابل توصیف بوده و
- این ویژگیها بصورت شرطی مستقل از یکدیگر باشند.
- تابع هدف f(x) بتواند هر مقداری را از مجموعه محدودv داشته باشد.

[۱] Bayes Optimal Classifier

تابع هدف $f: X \rightarrow V$ را در نظر بگیرید که در آن هر نمونه x توسط مجموعه ویژگی‌های $(a_1,\ldots a_n)$ مشخص می‌شود. حال صورت مسأله ای ناست که برای یک نمونه مشاهده شده مقدار تابع هدف یا بعبارت دیگر دسته بندی آنرا مشخص کنیم. در روش بیزی برای حل مسئله محتملترین مقدار هدف v_{map}محاسبه میشود:

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1,\ldots,a_n) \qquad (۱۰-۳)$$

این رابطه با استفاده از تئوری بیز بصورت زیر نوشته میشود :

$$\begin{aligned} v_{MAP} &= \arg\max_{v_j \in V} \frac{P(a_1,\ldots,a_n \mid v_j)P(v_j)}{P(a_1,\ldots,a_n)} \\ &= \arg\max_{v_j \in V} P(a_1,\ldots,a_n \mid v_j)P(v_j) \end{aligned} \qquad (۱۱-۳)$$

در رابطه فوق مقدار $P(v_j)$ با شمارش دفعاتی که v_j در مثالهای آموزشی مشاهده شده است محاسبه میشود. اما محاسبه $P(a_1,\ldots a_n \mid v_j)$ چندان عملی نیست مگر اینکه مجموعه داده آموزشی نسبتا بزرگی در دست باشد. روش یادگیری Naive Bayes Classifier بر پایه این فرض ساده (Naive) عمل میکند که: " مقادیر ویژگی‌ها بصورت شرطی مستقل هستند". در اینصورت برای یک مقدار هدف مشخص احتمال مشاهده ترکیب عطفی$(a_1,\ldots a_n)$ برابر است با حاصلضرب احتمال تک تک ویژگیها. در اینصورت رابطه‌ی فوق بصورت زیر در می‌آید[۴۶]:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j)\prod_{i=1}^{n} P(a_i \mid v_j) \qquad (۱۲-۳)$$

۳-۴-۲-۶- دسته‌بندی کننده‌ی Random Tree

پیش از هر توضیحی لازم است یکی از مهمترین دلایل انتخاب درخت تصمیم در این پروژه بیان شود، و آن اینکه درخت تصمیم بر خلاف بیشتر ساختارهای ماشین کلاس‌بندی قادر است با تعداد محدودی

نمونه از میان جامعه آماری آموزش خوبی بگیرد، البته با این شرط که نمونه‌های آموزشی بتوانند نماینده خوبی برای جامعه آماری خود باشند. البته در اواخر روال تحقیق پی بردیم که درخت تصمیم نمیتواند در این پروژه انتخاب خوبی باشد چرا که داده‌های مسئله‌ی مورد نظر پیچیده‌تر از آن هستند که بتوانند در این حجم کم (حدود ۸۰۰۰ نمونه) بتوانند تمام و یا بیشتر خصوصیات جامعه‌ی هدف را بیان کنند.

در هر یک از مسائل یادگیری ماشین با دو جنبه مختلف روبرو هستیم: یکی نحوه نمایش فرضیه ها و دیگری روشی که برای یادگیری برمی گزینیم [۴۷]. ساختار درخت تصمیم[۱] در یادگیری ماشین، یک مدل پیش‌بینی کننده می‌باشد که حقایق مشاهده شده در مورد یک پدیده را به استنتاج‌هایی در مورد مقدار هدف آن پدیده نگاشت می‌کند. تکنیک یادگیری ماشین برای استنتاج یک درخت تصمیم از داده‌ها، یادگیری درخت تصمیم نامیده می‌شود که یکی از رایج‌ترین روش‌های داده‌کاوی[۲] می‌باشد که ما بدین منظور از روش Random Tree استفاده خواهیم نمود.

علت نامگذاری این ساختار با درخت تصمیم این است که این درخت فرایند تصمیم گیری برای تعیین دسته یک مثال ورودی را نشان میدهد. ساختار درخت تصمیم به این شکل است که در آن نمونه‌ها را به نحوی دسته بندی میکند که از ریشه به سمت پائین رشد میکنند و در نهایت به گره های برگ میرسد، در ساختار این درخت داریم:

- هر گره داخلی یا غیر برگ با یک ویژگی مشخص می‌شود. این ویژگی سؤالی در رابطه با مثال ورودی مطرح میکند.
- درهر گره داخلی به تعداد جواب‌های ممکن با این سؤال شاخه[۳] وجود دارد که با مقدار آن جواب مشخص می‌شوند.

---

[۱] Desission Tree

[۲] Data Mining

[۱] Branch

- برگهای این درخت با یک کلاس و یا یک دسته از جوابها مشخص می‌شوند.

برای درک بهتر مفهوم درخت تصمیم از یک مثال استفاده می‌کنیم. فرض کنید می‌خواهیم نام‌های افراد را به دو دسته آقا و خانم کلاس‌بندی کنیم. بنابراین لازم است فهرستی از نام‌ها را در حالی که دارای برچسب f(خانم) و m(آقا) هستند را به عنوان ورودی در نظر بگیریم و با استفاده از آن مدل خود را به نحوی آموزش دهیم که بتواند نام جدید را بعنوان ورودی گرفته و مشخص نماید که این نام جدید در کدام‌یک از دو کلاس باید قرار گیرد.

جدول ۰-۱ مثالی از داده‌های یک مسأله‌ی کلاس‌بندی

نام	جنسیت
آرش	M
سارا	F
سینا	M
ایمان	M
زهرا	F

جدول ۰-۲ مثالی از مشخصه‌های استخراج شده از داده‌های یک مسأله

نام	اتمام با حرف صدادار	تعداد حروف صدادار	طول	جنسیت
آرش	۰	۱	۳	M
مینا	۱	۲	۴	F
صادق	۰	۱	۴	M
مهدی	۱	۱	۴	M
عاطفه	۰	۱	۵	F

حال بعنوان اولین گام باید بررسی کنیم که چه ویژگی‌هایی می‌توانند برای پی بردن به کلاس یک نام جدید مفید باشند. برخی از مثال‌ها عبارتند از طول، اولین یا آخرین کاراکتر، تعداد حروف صدادار، آیا با حرف صدادار تمام شده است و مثال‌هایی از این دست. بنابراین پس از گزینش ویژگی‌ها داده‌های ما به شکل زیر درمی‌آیند.

هدف ساخت یک درخت تصمیم است که در زیر یک نمونه را بعنوان مثال ملاحظه می‌کنید.

Length<۵

| length>۳

| | num-vowels<۲ m

| | num-vowels>=۲ f

| length<=۳ m

Length>=۵ f

به طور مشخص هر گره یک سنجش را بر روی یک ویژگی خاص انجام می‌دهد و ما بر اساس نتیجه این سنجش یکی از دو مسیر چپ و یا راست را انتخاب می‌کنیم و این عمل پیمایش درخت را تا جایی ادامه می‌دهیم که به یک گره‌ی برگ برسیم که در واقع در بردارنده برچسب کلاس مورد نظر یعنی f و یا m می‌باشد. بنابراین اگر نام جعفر را وارد این درخت کنیم در ابتدا با انجام مقایسه length<۵ به زیر درخت چپ هدایت می‌شویم و در آنجا نیز با انجام سنجش length>۳ مجدد به سمت راست هدایت می‌شویم و سپس با انجام مقایسه num-vowels<۲ به برگی با برچسب m می-رسیم.

در الگوریتم Random Tree در هر گره تعدادی از مشخصه‌ها که به صورت تصادفی انتخاب می‌شوند مورد ارزیابی قرار می‌گیرند. تعداد این مشخصه‌ها در این الگوریتم K نامیده می‌شود که یکی از پارامترهای ورودی الگوریتم بوده و باید توسط کاربر انتخاب گردد.

این الگوریتم در واقع نوع خاصی از الگوریتم Reduced Error Prouning (REPTree) می‌باشد [۴۸]. به طور کلی عمل هرس کردن[۱] در درختان به دو صورت پائین به بالا[۲] و یا بالا به پائین[۳] قابل انجام است که در هر یک از آن‌ها به ترتیب برگ‌ها و یا زیردرختان پیمایش و در صورت نیاز حذف می‌شوند. یکی از روش‌های رایج و ساده‌ی هرس Reduced Error Prouning نام دارد که یکی از روش‌های پائین به بالا بوده و در آن هر گره‌ی پیمایش شده با کلاسی که بیشترین شباهت را به داده‌ی موجود در آن گره دارد جایگزین می‌شود. پس از این جایگزینی میزان صحت عملکرد دسته‌بندی مورد ارزیابی مجدد قرار می‌گیرد و اگر صحت عملکرد کاهش نیابد این تغییر حفظ شده و در غیر اینصورت به حالت قبل بازگردانده می‌شود.

۳-۵- ترکیب نتایج دسته‌بندی کننده‌ها[۴]

ترکیب داده‌ها[۵] یکی از روش‌های رایج در دو مرحله‌ی داده‌کاوی می‌باشد، یکی در مرحله‌ی جمع آوری داده‌های اولیه که مورد بحث ما نخواهد بود، و دیگری در سطح دسته‌بندی کننده‌ها که نتایج آنها را با هم تلفیق نموده و به عبارت دیگر میان دسته‌بندهای مختلف رأی‌گیری می‌کند. در میان روش‌های مختلفی که برای ترکیب دسته‌بندی کننده‌ها وجود دارد، Voting یکی از انواع رایج می‌باشد که طرز کار آن دقیقا رأی‌گیری میان نظرات دو یا چند دسته‌بندی کننده می‌باشد. حال آنکه در چگونگی

---

[۱] Pruning
[۲] Bottom Up
[۳] Top Down
[۴] Classifier Ensembling
[۵] Data Fusion

انجام این رأی‌گیری می‌توان از روش‌های مختلفی استفاده نمود که یکی از آن‌هاMajority Voting یا رأی اکثریت نام دارد و مورد بحث ما نیز می‌باشد.

در مجموع سه روش رأی‌گیری را می‌توان نام برد که عبارتند از: Unanimity، Simple Majority و Plurality که در شکل ۳-۳ به صورت شماتیک نشان داده شده‌اند.

شکل ۳-۳ الگوهای مختلف رأی گیری در یک گروه ده نفره. در هر سه حالت تصمیم نهائی منطبق با رنگ مشکی خواهد بود. [۴۹]

فرض کنید خروجی‌های دسته‌بندی کننده‌های مختلف به صورت بردارهای باینری C بعدی می‌باشند:

$$[d_{i,1},\dots,d_{i,c}]^T \epsilon \{0,1\}^c, i = 1,\dots,L, where\ d_{i,j} = 1\ if\ D_i\ Labels\ x\ in\ \omega_j\ and\ 0\ otherwise \quad (۱۳-۳)$$

رأی‌گیری plurality منتج به یک ترکیب نتایج برای ω_j می‌گردد اگر:

$$\sum_{i=1}^{L} d_{i,k} = \max_{j=1}^{c} \sum_{i=1}^{L} d_{i,j} \quad (۱۴-۳)$$

این قاعده در اغلب موارد با نام Majority Vote خوانده می‌شود. همچنین این قاعده با مفهوم simple majority که به معنای ۵۰درصد بعلاوه‌ی یک می‌باشد نیز مطابقت دارد. در [۴۹] پیشنهاد می‌گردد

که برای plurality یک آستانه‌گذاری انجام شود و در مواردی که تنها دو دسته‌بندی کننده داریم نیز پیشنهاد می‌کند که یک کلاس سوم نیز بعنوان برچسب برای تمامی نمونه‌ها در نظر گرفته شود.

۳-۶- تفکیک داده‌های آموزش، تست و ارزیابی[۱]

در ایجاد یک ماشین دسته‌بندی کننده باید داده‌ها در سه مرحله به خدمت گرفته شوند، در مرحله‌ی اول نیاز است ماشین آموزش داده شود که بنابر توصیه‌ی برخی از منابع بهتر است ۷۰ درصد داده‌های در دسترس برای آموزش مورد استفاده قرار گیرد و از ۳۰ درصد باقیمانده برای مرحله‌ی دوم یعنی تست استفاده شود. در مرحله‌ی آموزش، ماشین در حال یادگیری می‌باشد و این بدان معناست که به ماشین گفته می‌شود هر داده‌ی ورودی باید چه برچسبی را داشته باشد و بنابراین ماشین، ساختار خود را به گونه‌ای طراحی می‌کند که بتواند این نگاشت را انجام دهد. در مرحله‌ی تست نیز داده‌های ورودی و خروجی به ماشین معرفی می‌شوند تا ماشین این موضوع را بررسی کند که با ساختار بدست آمده از مرحله‌ی آموزش با چه میزان خطائی مرحله‌ی تست را پشت سر می‌گذارد و سپس با اطلاعات بدست آمده از این مرحله به اصلاح ساختار خود می‌پردازد.

در مرحله‌ی ارزیابی نیز طراح و یا کاربر ماشین می‌تواند توان ماشین بدست آمده از دو مرحله‌ی قبل را بر روی داده‌هایی که ماشین تا کنون با آنها برخورد نکرده است، بررسی نماید.

۳-۶-۱- روش Cross Validation

همانطور که در بالا اشاره شد می‌توان داده‌ها را به دو قسمت آموزش و تست تقسیم نمود ،علاوه بر این تکنیک ساده، روش‌هایی نیز مورد استفاده هستند که می‌توان از آنها برای بهینه سازی عملکرد یک

[۱] Validation

دسته‌بندی کننده بهره گرفت که بدون توجه به آنها نمی‌توان قدرت واقعی یک دسته‌بندی کننده را در آزمایشات مورد نیاز برآورد نمود. یکی از پرکاربردترین این روش‌ها cross validation نام دارد که بر اساس آن مجموعه‌ی نمونه‌های آموزشی را به چند قسمت[1] مساوی تقسیم می‌کنند، این قسمت‌ها نباید با هم همپوشانی داشته باشند، تعداد این قسمت‌ها بسته به مسأله معمولا از میان اعداد ۳، ۵ و ۱۰ انتخاب می‌شود که البته محدودیتی برای این تعداد وجود ندارد و از لحاظ فنی میتوان حتی به تعداد نمونه‌ها نیز این قسمت‌بندی را انجام داد [۵۰]. الگوریتم این روش اینگونه است که اگر تعداد k قسمت در نظر گرفته شده باشد در k تکرار عمل کلاس‌بندی انجام می‌شود به اینصورت که در هر بار کلاس‌بندی یکی از k قسمت بعنوان داده‌ی تست و k-۱ قسمت دیگر برای آموزش در نظر گرفته می‌شوند و در نهایت دسته‌بندی کننده‌ای انتخاب می‌شود که در یکی از این k آزمایش نتیجه‌ی بهتری از k-۱ آزمایش دیگر داشته است.

[1] Fold

فصل چهارم

روش پیشنهادی

۴-۱- مقدمه

هدف از اجرای پروژه‌ی پیش رو چیزی به جز ایجاد یک ماشین دسته‌بندی‌کننده نیست که وظیفه‌ی آن نگاشت میان داده‌های پتروفیزیکی از یک سو و کلاس‌های شکستگی‌ها از سوی دیگر می‌باشد. نکته‌ی قابل توجه اینست که تا کنون چنین ماشینی با قابلیت اعتماد لازم در صنعت پا به عرصه‌ی حضور نگذاشته و این امر حاکی از پیچیدگی مسأله دارد.

آنچه دانستن آن در هنگام طراحی الگوریتم برای پروژه‌های داده‌کاوی مهم می‌باشد اینست که تا کنون بشر نتوانسته به یک فرمول واحد برای حل تمامی مسائل داده‌کاوی دست یابد و اینکه این کار اصلا با طبیعت مسأله‌ی داده‌کاوی مغایرت دارد. لذا در حل این مسائل باید دانش نسبتا گسترده و نیز تجربیات قبلی به خدمت گرفته شوند.

۴-۲- بررسی مراحل الگوریتم

داده‌هایی که در اختیار پروژه قرار گرفته‌اند شامل سه گروه داده به ازای هر چاه می‌باشند. دو گروه از داده‌ها شامل چاه‌نمودارهایی می‌باشد که از عمق‌های مختلف چاه برداشت شده‌اند و کاری که ما باید انجام می‌دادیم تلفیق این داده‌ها بوده است. به این صورت که با استناد به فیلد عمق باید سطرهای داده را در هر دو گروه پیدا کرده و آنها را در یک فایل قرار می‌دادیم که برای این‌کار یک برنامه به زبان جاوا توسعه داده شد. گروه دیگر از داده‌ها حاوی نظرات کارشناسان نفت مبنی بر تفسیر چاه‌نمودارهای تصویری می‌باشد که در واقع می‌توانیم بعنوان ناظر از آن استفاده کرده و این که هر عمق در چه کلاسی باید قرار بگیرد را می‌توانیم از روی داده‌های این فایل‌ها بدست آوریم.

پس از آماده‌سازی داده‌های اولیه آزمایشات زیادی بر روی آن‌ها انجام دادیم تا در آخر مراحل زیر برای حل مسأله طراحی گردیدند.

۴-۲-۱- پنجره‌گذاری

پس از آماده‌سازی داده‌های اولیه از چندین الگوریتم دسته‌بندی کننده از قبیل Baysian, SVM, KNN, C۴٫۵, Naïve Bayes, Random Forest و چند الگوریتم دیگر استفاده گردید که نتایج بدست آمده حاکی از عدم آماده بودن داده‌ها برای دسته‌بندی بود، چرا که دسته‌بندی در تمامی این الگوریتم‌ها شامل نتایج بسیار ضعیف بوده است. یک راه برای آماده نمودن داده‌ها استفاده از تکنیک پنجره‌گذاری می‌باشد که ما از آن بهره گرفتیم.

۴-۲-۱-۱- انتخاب تابع مورد استفاده بعنوان تابع تبدیل در پنجره‌گذاری

برای آنکه یک الگوریتم کارآمد بعنوان هسته‌ی الگوریتم پنجره‌گذاری انتخاب گردد بسیاری از توابعی که می‌توانند به عنوان خواص سیگنال به خدمت گرفته شوند از قبیل ضرائب موجک، گشتاورهای استاندارد ریاضی چهارگانه و ضرائب فوریه مورد استفاده قرار گرفتند و در نهایت گشتاور ریاضی دوم یعنی واریانس بعنوان تابع تبدیل انتخاب گردید. برای این انتخاب هر یک از توابع به خدمت گرفته شده و خروجی آن به عنوان داده جایگزین یکی از مقادیر داده در پنجره‌ی لغزان مورد استفاده قرار می‌گرفت. در گام بعدی سیگنال جدید را بعنوان مجموعه‌ی داده‌ها در نظر گرفته و عمل دسته‌بندی را با استفاده از الگوریتم‌های متعدد بر روی آن انجام دادیم. در آخر نیز با مقایسه‌ی نتایج این آزمایشات توانستیم واریانس را بعنوان بهترین انتخاب شناسایی نمائیم.

۴-۲-۱-۲- انتخاب طول مناسب پنجره

در ابتدا به یک تئوری برای کاهش آشفتگی سیگنال متوسل شدیم که بر اساس آن اگر یک سیگنال دارای گشتاور ریاضی چهارم کمتر باشد آنگاه دارای آشفتگی کمتری خواهد بود و معمولا این پارامتر

باعث افزایش کارائی الگوریتم‌های دسته‌بندی می‌شود. البته این تئوری همیشه به درستی عمل نمی-کند که با توجه به پیچیدگی‌های خاص داده‌های این پروژه در اینجا نیز به درستی عمل نکرد و ما ناگزیر به استفاده از تکنیک آزمون و خطا و مقایسه‌ی نتایج شدیم. در طی این گزینش در ابتدا طول پنجره را برابر ۳ در نظر گرفته و عمل پنجره گذاری، تبدیل سیگنال و در نهایت دسته‌بندی شد. این کار سپس با طول پنجره‌ی ۵، ۷، ۹،۱۱ و طول‌های بالاتر انجام دادیم و این کار تا آنجائی ادامه پیدا کرد که نتایج دسته‌بندی دچار کاهش عملکرد گردید و به این صورت مناسب‌ترین طول پنجره بدست آمد.

۴-۲-۲- نرمالسازی

پس از اینکه سیگنال‌های متشکل از مشخصه‌های نمونه داده‌ها را با استفاده از پنجره‌گذاری به یک سیگنال دیگر تبدیل نمودیم، نوبت به نرمال‌سازی این سیگنال‌های جدید می‌رسد که برای این‌کار از نرمالسازی میان ۰ و ۱ استفاده نمودیم. نکته‌ی قابل توجه در این قسمت اینست که نرمالسازی باید برای همه‌ی داده‌های تمام چاه‌ها به صورت همزمان اعمال گردد.

۴-۲-۳- انتخاب الگوریتم دسته‌بندی کننده

برای انتخاب الگوریتم مناسب برای دسته‌بندی از قواعد گفته شده در بخش ۳-۴-۱ استفاده گردید. بعنوان مثال با توجه به این قوانین می‌دانستیم که با توجه به عدم توازن کلاس‌ها نبایست از شبکه‌ی عصبی انتظار پاسخ مناسب را داشته باشیم و یا اینکه با توجه به پیچیدگی داده‌ها و عدم کفایت داده‌ها بعنوان یک نمونه آماری گویا از جامعه‌ی هدف، می‌دانستیم که الگوریتم‌های با بایاس بالا از قبیل Naïve Bayes می‌توانند انتخاب مناسبی باشند.

با این دانش قبلی بسیاری از الگوریتم‌های دسته‌بندی کننده مورد آزمون قرار گرفتند که در نهایت با بررسی دقیق نتایج، دو الگوریتم Naïve Bayes و Random Tree بعنوان مناسب‌ترین الگوریتم‌ها انتخاب گردیدند.

۴-۲-۴- ترکیب الگوریتم‌های دسته‌بندی کننده

در برخی از چاه‌ها مشاهده گردید که یک الگوریتم دسته‌بندی کننده از میان سه کلاس یکی را مثلا کلاس A را به خوبی تشخیص می‌دهد و برای دو کلاس دیگر عملکرد خوبی ندارد. این در حالی بود که یک الگوریتم دسته‌بندی کننده‌ی دیگر و یا همان الگوریتم اول ولی با پارامترهای دیگر در مورد یک کلاس دیگر مثلا کلاس B عملکرد خوبی از خود نشان می‌داد.

این مسأله باعث می‌شود ترکیب دسته‌بندی کننده‌ها از اهمیت خاصی برخوردار باشد و بنابراین از تکنیک رأی‌گیری اکثریت برای ترکیب نتایج دسته‌بندی کننده‌ها استفاده گردید.

۴-۲-۵- استفاده از Cross Validation

برای استفاده‌ی مؤثرتر از داده‌ها در مرحله‌ی آموزش از روش Cross Validation با Fold Number معادل ۱۰ استفاده گردید.

۴-۲-۶- روش ارزیابی نتایج

ما در ارزیابی‌های خود همواره از ماتریس درهم‌ریختگی[۱] بمنظور ثبت نتایج استفاده نمودیم. و در مقایسه‌های متعددی که باید انجام شود ضمن مقایسه‌ی این ماتریس‌های نتایج در برخی از قسمت‌ها از نمودار نیز استفاده گردیده.

این ماتریس اینگونه می‌باشد که هر سطر و هر ستون آن نماینده‌ی یک کلاس خواهد بود بعنوان مثال سطر دوم نماینده‌ی کلاس B ،ستون دوم نماینده‌ی کلاس B و نیز ستون اول نماینده‌ی کلاس A می‌باشد، حال اگر در المان سطر دوم و ستون اول عدد ۲۰ وجود داشته باشد بدین معنا خواهد بود که تعداد ۲۰ عدد از داده‌هایی که متعلق به کلاس B هستند به اشتباه بعنوان کلاس A دسته‌بندی شده-اند و اگر در المان سطر دوم و ستون دوم عدد ۳۰۰ را ملاحظه نمائید بدان معنا خواهد بود که تعداد ۳۰۰ المان که متعلق به کلاس B هستند به درستی تشخیص داده شده‌اند.

یک نکته‌ی حائز اهمیت در ارزیابی نتایج دسته‌بندی کننده‌ها اینست که در این پروژه اگر چه بالا بودن عملکرد کلی الگوریتم یک هدف می‌باشد ولی ایجاد توازن در میان کلاس‌ها نیز اصلی است که نباید از آن غافل شویم. بعنوان مثال اگر نتیجه‌ی دسته‌بندی دارای عملکرد ۹۰درصد باشد ولی تنها یکی از کلاس‌ها دارای صحت عملکرد بالای ۹۰ بوده و دو کلاس دیگر دارای صحت زیر ۵ درصد باشند این دسته‌بندی مناسب نخواهد بود. این درحالیست که اگر بعنوان مثال هر سه کلاس با صحت ۶۰درصد دسته‌بندی گردند این نتیجه بسیار خوشآیند و قابل قبول برای بخش صنعت خواهد بود.

فرض کنید داده‌های X چاه در اختیار پروژه قرار دارد، برای اثبات صحت عملکرد هر الگوریتم به این صورت عمل می‌کنیم که از داده‌های X-۱ چاه برای آموزش استفاده کرده و با استفاده از ماشین ایجاد شده داده‌های یک چاه باقیمانده (که در مرحله‌ی آموزش حضور نداشت) را دسته‌بندی کرده و ماتریس درهم‌ریختگی را برای ارزیابی صحت عملکرد آن تشکیل می‌دهیم.

[۱] Confusion Matrix

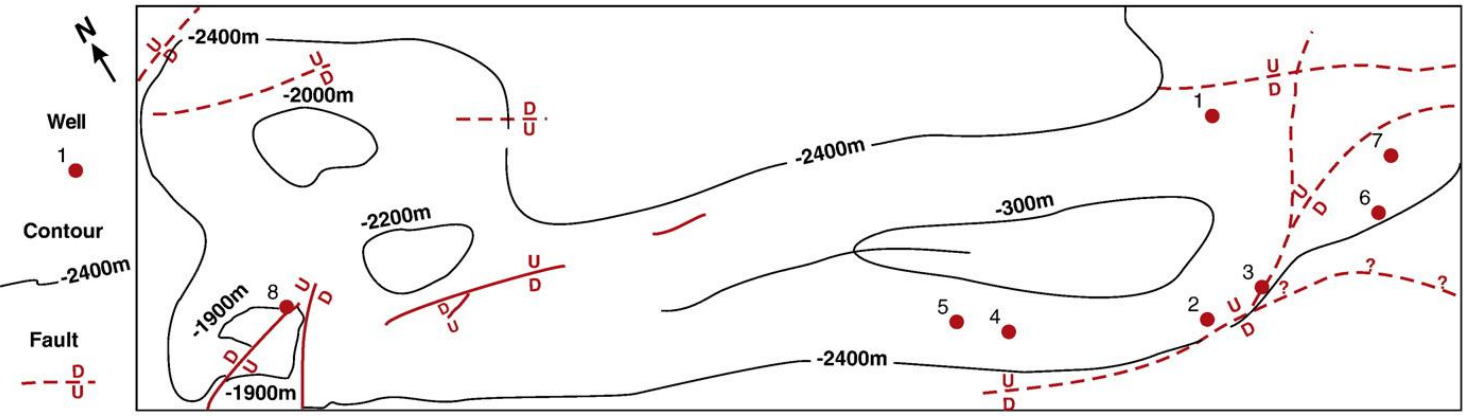
فصل پنجم

معرفی داده‌های تحقیق و

میدان مورد مطالعه

۵-۱- داده‌های مسأله

برای انجام این تحقیق داده‌های مربوط به هشت حلقه چاه از میدان گچساران در اختیار قرار گرفت، در تصویر ۵-۱ میتوانید نقشه منحنی میزان زیرسطحی[۱] بخش بالایی سازند آسماری این میدان را ملاحظه نمائید که در آن موقعیت هشت حلقه چاه مذکور نیز تصویر شده است.



شکل ۵-۱ نقشه منحنی میزان زیرسطحی تهیه شده از میدان نفتی گچساران و موقعیت چاه‌های مورد مطالعه [۶]

اطلاعاتی که از این چاه‌ها در دست داریم شامل دو دسته می‌باشند یکی چاه‌نمودارهای پتروفیزیکی که عموما به آنها نمودار نیز اطلاق می‌شود و در ادامه به تفصیل مورد بررسی قرار می‌گیرد، و دیگری تفسیر چاه‌نمودارهای تصویری که به عنوان مرجعی برای شناسایی شکستگی‌ها و ویژگی‌های آنها مورد استفاده قرار می‌گیرند. بنابراین باید به دنبال روشی باشیم تا بتوانیم با استفاده از آن، از روی چاه-نمودارهای پتروفیزیکی به همان نتایجی دست یابیم که تفسیر چاه‌نمودارهای تصویری بیان می‌کنند. نتیجه کار در نهایت یک ماشین دسته‌بندی کننده[۲] خواهد بود که در طول روند آموزش آن باید از تفسیر چاه نمودارهای تصویری به عنوان ناظر استفاده نمائیم.

[۱] UGC Map
[۲] Classifier

یک نکته که در مرحله گزینش داده‌ها باید به آن توجه شود اینست که تمامی چاه‌نمودارها می‌بایست مربوط به اعماقی باشد که متعلق به سازند آسماری هستند و نباید لاگ سازند فوقانی یعنی گچساران و یا سازند پائینی یعنی پابده در میان داده‌های مورد مطالعه جای گیرد، برای رسیدن به این هدف باید چاه‌نمودارهای متعلق به سازندهای بالایی و پائینی را از میان داده‌ها حذف نمائیم. بدین منظور در تحقیق پیش رو به اطلاعات موجود در جدول ۵-۱ استناد شده است که از [۲۶] استخراج گردیده است.

جدول ۵-۱ اطلاعات عمقی مربوط به چاه‌های آسماری پیش و پس از تصحیحات عمقی [۲۶]

شماره چاه	*Zone A*	پابده	*TD*	اطلاعات چاه‌نمودارها		اطلاعات چاه‌نمودارها پس از تصحیح بر اساس سازند آسماری		اختلاف ضخامت در ابتدای چاه (متر)	اختلاف ضخامت در انتهای چاه (متر)
	A۱			عمق آغازین	عمق انتهایی	عمق آغازین	عمق انتهایی		
۱	۲۳۷۳	-	۲۷۰۲	۲۳۷۳,۳۲۵۲	۲۶۹۸,۵۴۶۸	۲۳۷۳,۳۲۵۲	۲۶۹۸,۵۴۶۸	۰,۳۲۵۲	-۳,۴۵
۲	۲۴۹۲	-	۲۹۸۰	۲۴۹۲,۶۵۴۴	۲۹۶۹,۰۵۶۸	۲۴۹۲,۶۵۴۴	۲۹۶۹,۰۵۶۸	۰,۶۵۴	-۱۰,۹
۳	۲۱۴۲	-	۲۸۰۸	۲۱۴۲,۵۸۹۷	۲۷۸۹,۹۸۴۲	۲۱۴۵,۹۴۲۵	۲۷۷۸,۵۵۴۲	۳,۹۴	-۲۹,۴۴
۵	۱۷۸۵	۲۲۸۰	۲۶۴۳	۱۷۸۵,۰۵۹۵	۲۶۱۵,۹۴۳۳	۱۷۸۵,۰۵۹۵	۲۲۷۹,۹۰۱۷	۰,۰۵۶	-۰,۰۹۸
۶	۲۲۸۲	۲۷۲۰	۲۷۵۰	۲۲۸۲,۱۹	۲۷۲۹,۳۳۱۶	۲۲۸۲,۱۹	۲۷۱۹,۸۸۲۸	۰,۱۹	-۰,۱۱۷۲
۷	۲۱۱۹	۲۳۸۰	۲۴۱۰	۲۱۱۹,۱۲۲۱	۲۳۷۹,۸۷۸۴	۲۱۱۹,۱۲۲۱	۲۳۷۹,۸۷۸۴	۰,۱۲۲۱	-۰,۱۲۱۶
۸	۲۳۷۰,۵	-	۲۷۲۶	۲۳۷۰,۶۴۲۸	۲۷۲۲,۸۳۵	۲۳۷۰,۶۴۲۸	۲۶۸۴,۲۸۱۷	۰,۱۴۲۸	-۴۱,۷۱

۵-۲- کلاس‌های شکستگی

همانطور که در فصل اول گفته شد هدف نهایی این تحقیق دسته بندی داده‌ها بر اساس نوع شکستگی می‌باشد بعبارت دیگر در هر عمق باید تعیین شود که دیواره‌ی چاه در این عمق آیا فاقد شکستگی است، دارای شکستگی بسته شده است و یا آنکه دارای شکستگی باز می‌باشد. پس در واقع فضای مسئله دارای سه کلاس زیر خواهد بود:

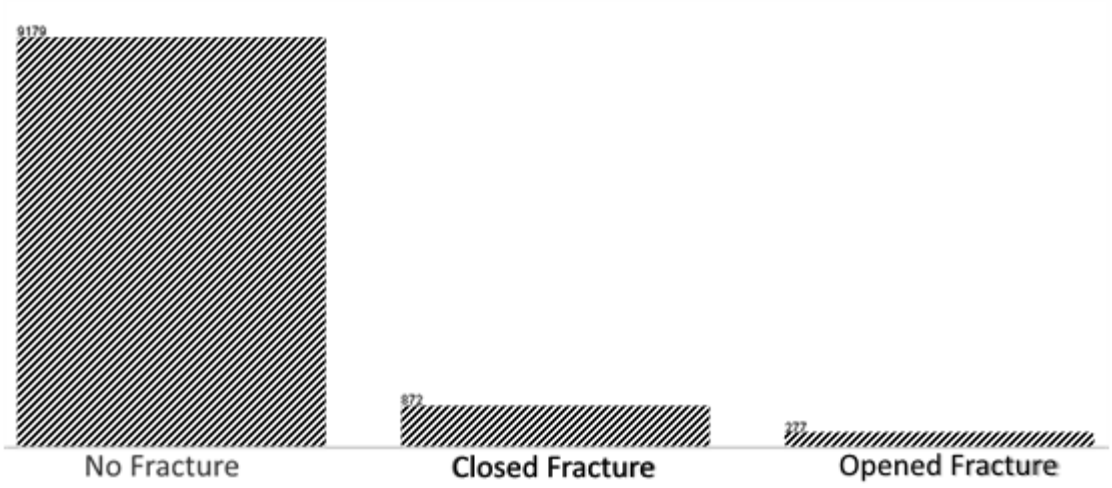
۱. کلاس مربوط به نواحی فاقد شکستگی که ما از این پس آن را با برچسب no-frac نشان خواهیم داد.
۲. کلاس مربوط به نواحی دارای شکستگی بسته که ما از این پس آن را با برچسب closed نمایش می‌دهیم.
۳. کلاس مربوط به نواحی دارای شکستگی باز که ما از این پس آن را با برچسب opened نشام می‌دهیم.

برای آنکه یک دید کلی از نحوه توزیع شکستگی‌ها داشته باشید به جدول ۵-۲ توجه نمائید که با توجه به تفاوت فاحش میان تعداد شکستگی‌ها در این چاه‌ها، اطلاعات این جدول خود می‌تواند بیانگر پیچیدگی رفتاری شکستگی‌ها و اهمیت انجام این پروژه باشد.

جدول ۵-۲ تعداد شکستگی‌های باز و بسته در چاه‌های مورد مطالعه

شماره چاه	تعداد کل داده‌ها	تعداد شکستگی‌های باز	تعداد شکستگی‌های بسته
۱	۲۲۸۹	۲۵	۶۴۷
۵	۴۰۴۲	۲۱	۹۰
۷	۱۷۳۳	۴۴	۱۳۲
۸	۲۳۴۹	۱۸۷	۳

در شکل ۵-۲ نیز میزان پراکندگی سه کلاس در مجموعه‌ی داده‌های چهار چاه ۱، ۵، ۷ و ۸ تصویر شده است.



شکل ۵-۲ پراکندگی کلاس‌های شکستگی در مجموع داده‌های چهار چاه ۱، ۵، ۷ و ۸

۵-۳- انتخاب مشخصه‌ها

گام دوم که یکی از مهمترین گام‌های این فصل و حتی کل پروژه محسوب می‌شود یافتن چاه-نمودارهای مشترک میان چاه‌های مورد مطالعه است. که ما برای سنجش میزان کارائی دو روش مختلف یک بار سعی در به حداکثر رساندن مشخصه‌ها خواهیم داشت که به تبع انجام این کار تعداد نمونه‌ها به حداقل ممکن کاهش می‌یابد. در دسته‌ی دیگری از آزمایشات تعداد نمونه‌ها را به حداکثر خواهیم رسانید که در نتیجه‌ی آن تعداد مشخصه‌های مشترک میان چاه‌ها کاهش چشم‌گیری خواهد داشت.

از آنجایی که داده‌های تحت اختیار قرارگرفته، شامل سه دسته داده بودند قبل از هر اقدامی برای گزینش چاه‌نمودارها باید داده‌ها را با هم ادغام می‌کردیم. این سه دسته داده عبارتند از:

- داده‌های حاصل از تفسیر چاه‌نمودارهای تصویری
- داده‌های مربوط به چاه‌نمودارهای پتروفیزیکی اندازه‌گیری شده
- داده‌های مربوط به چاه‌نمودارهای پتروفیزیکی تفسیری

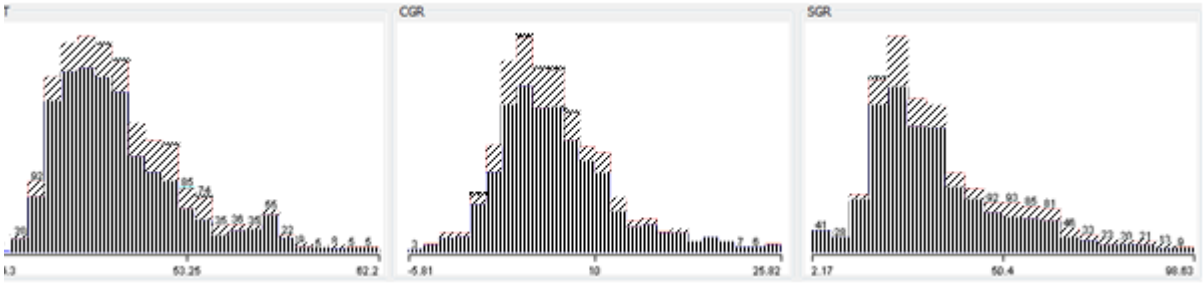
داده‌های حاصل از تفسیر چاه‌نموارهای تصویری در واقع حکم ناظر را دارند که با استفاده از آنها تعیین می‌کنیم که یک عمق خاص در چه کلاسی قرار می‌گیرد. داده‌های پتروفیزیکی اندازه‌گیری شده و داده‌های پتروفیزیکی تفسیری در مجموع داده‌های پتروفیزیکی نامیده می‌شوند. بنابراین ما باید این سه دسته داده را به صورت متناظر در کنار یکدیگر قرار دهیم و یک دسته داده ایجاد نمائیم. برای ادغام این داده‌ها یک برنامه به زبان جاوا ایجاد گردید که دارای ورودی و خروجی از نوع CSV می‌باشد.

جدول ۵-۳ چاه‌نمودارهای موجود در هر چاه

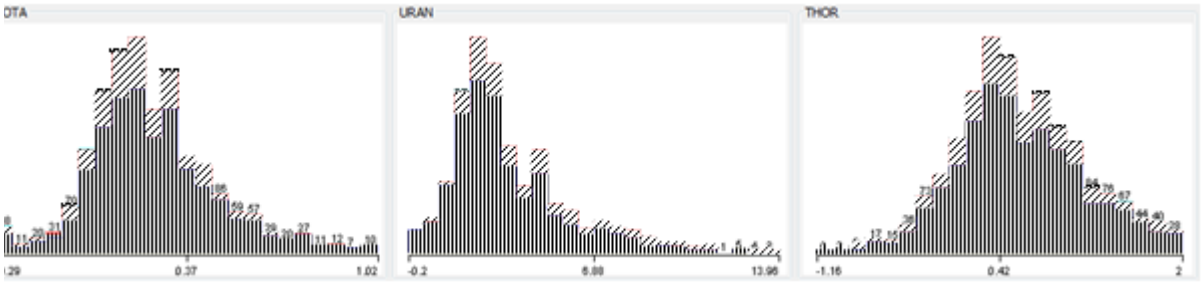
چاه \ ویژگی	CALI	DT	CGR	SGR	RHOB	NPHI	RT	FLAG	PEF	GR	POTA	URAN	THOR	PHIE	SWE	CALCITE	DOLOMITE	ANHYDRITE	SHALE	ILM	ILD	MSFL	LLD	LLS
۱	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
۲	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
۳	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
۴	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
۵	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗
۶	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
۷	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗
۸	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	✗

از میان این هشت حلقه چاه چهار تای آن دارای یازده چاه‌نمودار قابل استفاده در داده‌کاوی هستند چرا که این یازده چاه‌نمودار در چهار چاه مذکور حضور دارند.

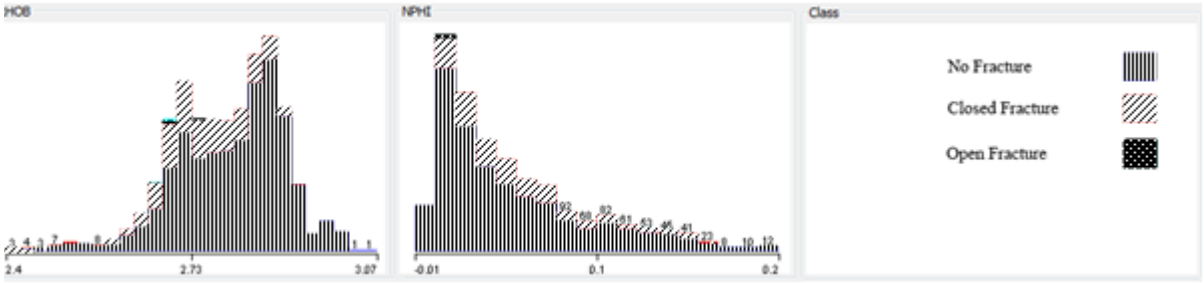
جدول ۵-۴ تفکیک نواحی شکسته‌ی باز، شکسته‌ی بسته و شکسته‌نشده با استفاده از هیستوگرام داده‌های پتروفیزیکی در چاه شماره ۱



الف) DT ب) CGR پ) SGR



ج) POTA چ) URAN ح) THOR



خ) RHOB د) NPHI ذ) Class Signatures

این یازده چاه‌نمودار عبارتند از Caliper, Gamma ray (SGR, CGR), uranium, thorium, potassium, sonic(DT), resistivity (RT), density (PEF , RHOB), neutron (NPHI) ، در

ضمن تفسیر چاه‌نمودارهای تصویری هر چاه نیز در اختیار است و همانطور که اشاره شد ما از آن بعنوان مرجعی برای شناسایی کلاس هر زون استفاده خواهیم نمود.

در شکل ۵-۳ بعنوان نمونه هیستوگرام ۸ چاه‌نمودار از بین یازده چاه‌نمودار مورد نظر در چاه اول نشان داده شده‌اند، با نگاه به هیستوگرام‌های چاه اول متوجه می‌شویم که در نواحی با شکستگی بسته، مقدار RHOB, THOR اندکی افزایش می‌یابد و نیز در این نواحی مقدار POTA, URAN, NPHI, SGR, CGR , DT, CALI, PEF اندکی کاهش می‌یابد. همچنین برای شکستگی‌های باز نیز می‌توان پی برد که مقدار RHOB, THOR با افزایش همراه است در حالی‌که مقدار URAN, DT , PEF, POTA, SGR کاهش داشته‌اند. توجه داشته باشید که این نتایج با توجه به هیستوگرام داده‌های پتروفیزیکی مربوط به چاه شماره ۱ بدست آمده‌اند.

هیستوگرام‌های معرفی شده در شکل ۵-۳ نمایی یک بعدی از ارتباط میان هر داده و سه کلاس مورد نظر ما می‌باشد حال اگر داده‌های فوق را به صورت دوبعدی نمایش دهیم قدرت تفکیک ما افزایش می‌یابد هر چند پردازش ممکن است پیچیده‌تر شود. همچنین اگر داده‌ها را به صورت سه بعدی تحلیل کنیم باز هم قدرت تفکیک کلاس‌ها افزایش و پیچیدگی افزایش می‌یابد. از آنجا که در این پروژه میزان دقت بسیار اهمیت دارد تصمیم بهتر آن است که از تمامی ابعاد موجود که در اینجا یازده بعد می‌باشد استفاده نمائیم.

در نهایت ما در فصل بررسی نتایج به منظور انجام عملیات دسته‌بندی، یک بار از داده‌های چهار چاه و بار دیگر از داده‌های هفت چاه استفاده خواهیم نمود چرا که همانطور که پیش‌تر نیز ذکر شد اگر زیاد بودن تعداد مشخصه‌ها مد نظر قرار گیرد آنگاه یازده مشخصه‌ی مهم و مشترک تنها میان چهار چاه وجود دارند.

۵-۴- داده‌های فاقد مقدار (تهی)

داده‌هایی که دارای مقدار مقدار (۹۹۹,۲۵-) می‌باشند در واقع *null* هستند. در چاه شماره‌ی ۱ که مجموعا دارای ۲۲۸۹ داده می‌باشد تعداد ۸۷ داده در مشخصه‌های *Anhydrite, Shale, RT, Dolomite* دارای مقدار *null* هستند. در این میان ۸۵ داده در صدر چاه قرار دارند و دو داده نیز در میانه‌های چاه. در چاه شماره‌ی ۵ نیز تعداد پنج داده در میانه‌های چاه قرار دارند که فاقد مقدار در مشخصه‌ی SGR می‌باشند.

برای مقابله با این داده‌های فاقد مقدار سه راه می‌توانند پیشنهاد شوند، یکی حذف این داده‌ها، دیگری درونیابی و برون‌یابی این داده‌ها و آخرین روش، نادیده گرفتن این نکته که با داده‌های فاقد مقدار مواجه هستیم می‌باشد.

روش اول از لحاظ مفهموی قابل پذیرش نیست چرا که با توجه به اینکه ما قصد داریم در مراحل بعد از روش پنجره گذاری استفاده کنیم بنابراین باید داده‌ها در شرایط یکسان و با فاصله‌ی یکسان از یکدیگر برداشت شده باشند، که در حال حاضر این اصل محقق می‌باشد و ما نباید با حذف تعدادی از داده‌ها آن را بر هم بزنیم. نادیده گرفتن داده‌های فاقد مقدار و یا جایگزین نمودن آنها با مقداری مثل ۹۹۹- نیز ادامه‌ی عملیات را دچار خطای زیاد خواهد نمود. پس بهترین راهی که پیشنهاد می‌شود استفاده از درونیابی می‌باشد که ما در اینجا از روش درونیابی Cubic Spline Interpolation استفاده نموده‌ایم. در مجموع تعداد پانزده مقدار نیاز به درون‌یابی داشتند. در مورد داده‌هایی که فاقد مقدار بوده و در ابتدای چاه شماره‌ی ۱ قرار دارند با توجه به اینکه این داده‌ها متعلق به نواحی فاقد شکستگی هستند دارای اهمیت کمتری بوده، بعلاوه‌ی اینکه از آنجا که تعداد هشتاد و پنج داده در کنار هم قرار گرفته‌اند لذا برون‌یابی آن‌ها با خطلای بسیار قابل ملاحظه‌ای همراه می‌باشد، لذا این هشتادوپنج داده حذف گردیدند.

فصل ششم

نتایج تحقیق و بحث

۶-۱- مقدمه

همانگونه که پیش‌تر در فصل سوم بیان شد، ما با دو نوع از چاه‌نمودارهای پتروفیزیکی مواجه هستیم که عبارتند از چاه‌نمودارهای اندازه‌گیری شده و چاه‌نمودارهای تفسیری. از آنجایی که چاه-نمودارهای تفسیری در تعداد کمتری از چاه‌ها حضور داشتند، علیرغم آگاهی از اهمیت آنها ما یکبار روش پیشنهادی را بدون در نظر گرفتن این نوع چاه نمودارها انجام دادیم و بار دیگر همین آزمایشات را بر روی تعداد کمتری چاه ولی با حضور چاه‌نمودارهای تفسیری اجرا نمودیم تا بتوانیم قدرت کلاس‌بندی روش پیشنهادی را در هر دو شرایط بسنجیم.

قبل از آغاز آزمایشات لازم است اشاره‌ای داشته باشیم به این نکته که در این تحقیق کسب اطمینان از قابلیت تعمیم[۱] جزء جدایی ناپذیر آزمایشات می‌باشد چرا که هدف نهایی ما در استفاده از داده‌های چاه‌های مختلف بدست آوردن مدلیست که بتواند بر روی چاه‌هایی عمل نماید که اطلاعاتی در مورد شکستگی‌های آنها در دست نیست.

۶-۲- چگونگی سنجش میزان کارائی دسته‌بندی

همانگونه که در قسمت ۳-۲ مطرح گردید اغلب داده‌ها در یکی از سه کلاس یعنی در کلاس no-frac قرار می‌گیرند و بنابراین اگر حتی بدون انجام عمل دسته‌بندی همه‌ی داده‌ها را در این کلاس قرار دهیم آنگاه به دقتی بالاتر از ۹۰٪ دست یافته‌ایم، از طرف دیگر از نقطه‌نظر مهندسی اکتشاف آنچه در این تحقیق اهمیت دارد شناسایی صحیح نواحی شکسته می‌باشد. بنابراین بعنوان مثال اگر به فرآیندی دست یابیم که در آن نواحی شکسته با دقت بیش از ۴۰ درصد و نواحی فاقد شکستگی با دقت بالاتر از ۶۰ درصد شناسایی شوند آنگاه گامی موثر برای بهره‌برداری صنعتی از این تحقیق برداشته‌ایم. در

[۱] Generalization

یک جمله می‌توان گفت که دست یافتن به توازن هر چه بیشتر در میزان دقت همه‌ی کلاس‌ها بسیار مطلوب می‌باشد حتی اگر باعث کاهش میزان دقت کل گردد.

البته بدیهی است که دقت کل نباید از حد مشخصی کمتر شود. بعنوان مثال دردسته‌بندی داده‌ها به دو کلاس، دقت زیر ۵۰ درصد در عمل به معنای این است که هیچ کاری انجام نشده، چرا که دسته-بندی به صورت تصادفی دارای احتمال پاسخ ۵۰ درصد می‌باشد.

بنابراین ما در طی تمامی آزمایشات میزان کارائی روش‌های مختلف را بر اساس کارائی آن روش در شناسایی هر یک از سه کلاس و نیز میزان کارائی در شناسایی کل داده‌ها مورد سنجش قرار می‌دهیم.

۶-۳- انتخاب حداکثری تعداد مشخصه‌ها

همانگونه که در مقدمه‌ی این فصل گفته شد در این قسمت از آزمایشات تمرکز بر روی به خدمت گرفتن تعداد حداکثری چاه‌نمودارها به منظور دسته‌بندی کارامدتر می‌باشد هر چند می‌دانیم که این کار باعث می‌شود داده‌های آزمایش به حداقل ممکن کاهش یابند.

۶-۳-۱- گزینش مشخصه‌ها و داده‌ها

از میان هشت حلقه چاه مورد نظر چهار تای آن دارای یازده چاه‌نمودار قابل استفاده در داده‌کاوی هستند چرا که این یازده چاه‌نمودار در چهار چاه مذکور حضور دارند. این یازده چاه‌نمودار عبارتند از Caliper, Gamma ray (SGR, CGR), uranium, thorium, potassium, sonic(DT), resistivity (RT), density (PEF , RHOB), neutron (NPHI) ، در ضمن تفسیر چاه‌نمودارهای تصویری هر چاه

نیز در اختیار است و همانطور که اشاره شد ما از آن بعنوان مرجعی برای شناسایی کلاس هر زون استفاده خواهیم نمود.

در جدول ۶-۱ مشخصه‌ها و چاه‌های انتخاب شده برای این قسمت از تحقیق با رنگ متفاوت نشان داده شده‌اند.

جدول ۶-۱ چاه‌نمودارهای موجود در هر چاه

چاه / ویژگی	CALI	DT	CGR	SGR	RHOB	NPHI	RT	FLAG	PEF	GR	POTA	URAN	THOR	PHIE	SWE	CALCITE	DOLOMITE	ANHYDRITE	SHALE	ILM	ILD	MSFL	LLD	LLS
۱	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
۲	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
۳	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
۴	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
۵	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗
۶	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
۷	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗
۸	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	✗

۶-۳-۲- انتخاب الگوریتم دسته‌بندی

پس از انجام آزمایشات لازم که به منظور دسته‌بندی داده‌ها در سه کلاس و با کمک تعداد قابل توجهی از الگوریتم‌های دسته‌بند انجام دادیم به این نتیجه ناامیدکننده رسیدیم که در داده‌های مسئله، یک رابطه‌ی قابل درک برای الگوریتم‌های کلاسه‌بند وجود ندارد. الگوریتم‌هایی که در این آزمایشات مورد بررسی قرار گرفتند عبارتند از: الگوریتم درخت استنتاج C۴,۵، شبکه‌ی عصبی چند لایه، شبکه‌ی عصبی، بیزین، نزدیکترین همسایگی‌ها، Random Tree و همچنین الگوریتم‌های ترکیب اطلاعات از قبیل الگوریتم بگینگ و بوستینگ و رای‌گیری. در میان تمامی الگوریتم‌های دسته‌بندی الگوریتم Naïve Bayes و Random Tree نتایج بهتری را در بر داشتند که در ادامه‌ی آزمایشات بر روی این دو الگوریتم متمرکز خواهیم شد.

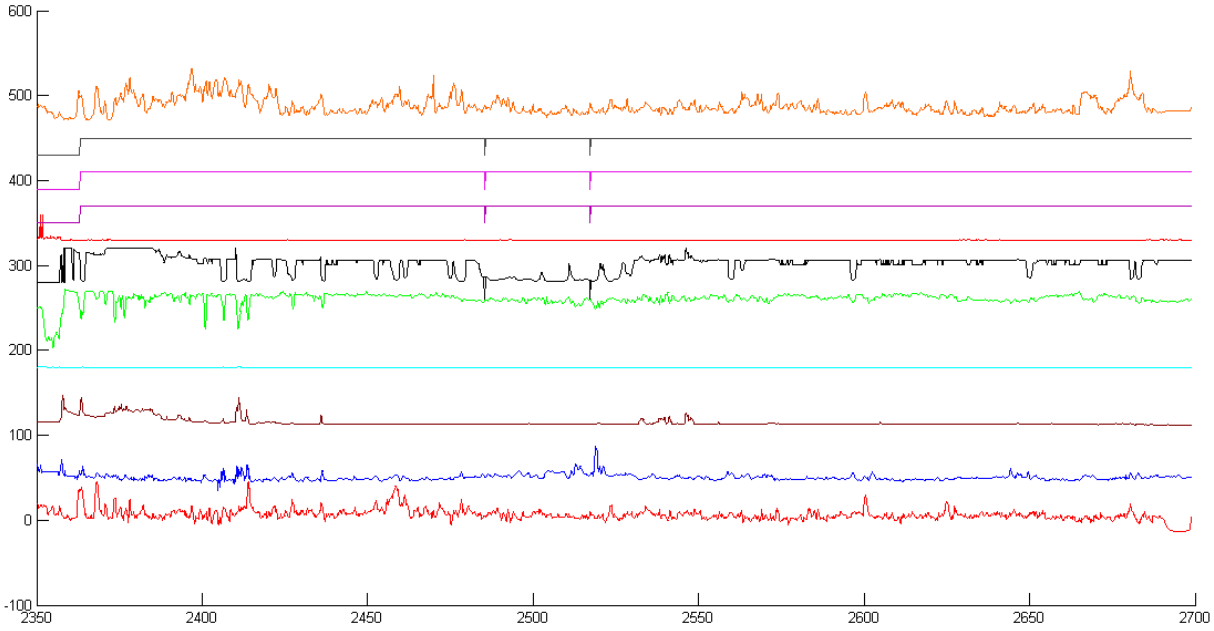
برای مرحله پیش‌پردازش دو گام اصلی خواهیم داشت. یکی ایجاد مشخصه‌های جدید و دیگری نرمال-سازی داده‌ها و شاید هم نیاز به حذف داده‌های پرت داشته باشیم که در ادامه مشخص خواهد شد.

۶-۳-۳- ایجاد مشخصه‌های جدید با استفاده از تکنیک پنجره‌گذاری

همانطور که در قسمت ۴-۲-۱ مطرح گردید هیچ یک از الگوریتم‌های دسته بندی نتوانستند به نحو قابل قبولی دسته‌بندی داده‌ها را انجام دهند بنابراین ما باید به دنبال انجام یکی از روشهای مطرح در پیش‌پردازش باشیم که همانا استخراج مشخصه‌های جدید از روی مشخصه‌های داده‌های خام می‌باشد. در این قسمت قصد داریم به جای داده‌های خام از داده‌های حاصل از پنجره‌گذاری استفاده نمائیم.

نکته‌ی حائز اهمیت در داده‌های مسئله اینست که هر یک از مشخصه‌ها (لاگ‌ها) در واقع یک سیگنال با محوریت عمق چاه می‌باشد. برای روشن شدن مطلب به شکل ۶-۱ توجه فرمائید که نشان‌دهنده‌ی تعداد یازده سیگنال می‌باشد و هر یک از این سیگنال‌ها متعلق به یکی از لاگ‌های چاه شماره ۱ می-

باشد. این تصویر فقط به منظور نمایش رفتار سیگنال‌ها ترسیم شده است و بنابراین مقادیر هر سیگنال با تغییر اندازه و مقداری شیفت به نحوی تنظیم شده‌اند که برروی یکدیگر قرار نگیرند.



شکل ۶-۱ رفتار سیگنال‌های به دست آمده از یازده لاگ متعلق به چاه شماره ۱

با مقدمه‌ی بیان شده‌ی بالا ایده اولیه‌ای که به ذهن می‌رسد اینست که رفتار سیگنال، خود می‌تواند بعنوان منشأی برای شناسایی مشخصه‌های جدید در نظر گرفته شود. بنابر این در ادامه این قسمت قصد داریم با بخش‌بندی[۱] هر یک از سیگنال‌ها به شناسایی مشخصه‌های مبین رفتار سیگنال در هر بخش[۲] بپردازیم. به طور طبیعی پیش‌بینی می‌کردیم انتخاب طول مناسب پنجره‌ی بخش‌بندی، میزان همپوشانی پنجره، انتخاب الگوریتم‌های مناسب به منظور استخراج ویژگی‌های سیگنال و در نهایت کاهش ابعاد ویژگی‌های استخراج شده، می‌توانند در این مرحله مخاطره برانگیز باشند.

الف. انتخاب یک فرمول مناسب برای محاسبه‌ی ویژگی نمودار در هر پنجره

[۱] Segmentation
[۲] Segment

طی چندین مرحله آزمایش که لحاظ نمودن نتایج رقمی آن‌ها خارج از حجم این مستند می‌باشد انواع فرمول‌هایی که می‌توان از آن‌ها برای استخراج مشخصه‌ی سیگنال در هر پنجره استفاده نمود مورد بررسی قرار گرفتند که عبارت‌اند از گشتاورهای‌ریاضی‌استاندارد[۱] چهارگانه، ضرایب فوریه[۲] و ضرایب تبدیل موجک[۳].

در پایان این نتیجه حاصل گردید که استفاده از گشتاور ریاضی دوم یعنی واریانس[۴] بهترین انتخاب برای محاسبه‌ی ویژگی نمودار در هر پنجره می‌باشد.

ب. انتخاب طول پنجره‌ی مناسب

برای ورود به مبحث بخش‌بندی سیگنال در ابتدا طول بخش‌ها را ثابت در نظر می‌گیریم ولی این امکان را نیز از ذهنمان دور نمی‌کنیم که شاید در ادامه کار با نتایج نامطلوبی برخورد کرده و مجبور به استفاده از طول پنجره متغیر شویم. باز هم برای ساده‌تر شدن کار در اینجا میزان همپوشانی پنجره‌ها را k-۱ در نظر می‌گیریم که در آن k بیانگر طول پنجره می‌باشد.

در این مرحله از آزمایشات یک برنامه به زبان متلب[۵] توسعه داده شد که بعنوان ورودی می‌بایست آدرس فایل CSVی حاوی داده‌ها و نیز طول پنجره را بگیرد و در پایان به ازای هر یک از داده‌ها یک مشخصه به نام ‘X_S’ را در فایل CSV وارد می‌کند. منظور از X در اینجا نام مشخصه‌ی مورد نظر و منظور از Sهمان انحراف از معیار می‌باشد. پس باز هم برای ساده‌سازی در اولین آزمایش از یک روش

[۱] Standardized Moments
[۲] Fourier
[۳] Wavelet
[۴] Variance
[۵] Matlab

بسیار ساده که همان تابع توزیع احتمال نرمال می‌باشد بهره گرفتیم. لازم به ذکر است که قبل از انجام عملیات پنجره‌گذاری، نرمالسازی بر روی داده‌ها اعمال نشده است.

بنابر توضیحات بالا قصد ما در این مرحله از کار اینست که بتوانیم واریانس یک قطعه از نمودار را محاسبه نموده و آن را بعنوان یک ویژگی به نمونه‌های موجود در آن قطعه سیگنال نسبت دهیم. اما سؤالی که در اینجا مطرح می‌شود این است که طول قطعات باید به چه اندازه باشد تا بهترین نتیجه در مرحله طبقه‌بندی را به همراه داشته باشد. پاسخ این سؤال این خواهد بود که ما از اندازه ایده‌آل بی اطلاع بوده و این عدد کاملا وابسته به داده‌های مسأله می‌باشد. پس یک راه استفاده از چندین طول متغیر و در نهایت استفاده از یک الگوریتم انتخاب ویژگی بوده و راه دیگر استفاده از یک تابع بهینه ساز می‌باشد.

ما برای انتخاب طول پنجره‌ی مناسب به این شکل عمل نمودیم که در ابتدا از داده‌های خام در فرآیند دسته‌بندی استفاده نمودیم و سپس از داده‌های بدست آمده از پنجره‌گذاری با طول ۳، ۵، ۷، ۹، ۱۱، ۱۳، ۱۵، ۱۷و ۱۹ استفاده نمودیم. در هر یک از این مراحل میزان کارائی فرآیند دسته بندی بر روی چاه‌های مختلف بررسی و مقایسه گردید. که در نتیجه طول پنجره‌ی ۱۱ نسبت به سایر طول‌ها عملکرد بهتری را نشان داد به نحوی که پنجره‌های با طول بالاتر به ترتیب دارای عملکردهای ضعیف-تری نسبت به طول پنجره‌ی انتخاب شده بودند.

در زیر نتیجه‌ی آزمایشات بر طبق فرض‌های قسمت ۴-۲ و با استفاده از الگوریتم Naïve Bayes آورده شده است:

جدول ۰-۲ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۳

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱۲٦٤ ۱۸٦ ۸۲ | a = no-frac ٥۱۳ ۹٥ ۳۹ | b = closed ۲٥ ۰ ۰ | c = opened	a b c <-- classified as ۱۸۳۸ ۱۹۷۰ ۱۲۳ | a = no-frac ٤۳ ٤۷ ۰ | b = closed ۷ ۱۲ ۲ | c = opened	a b c <-- classified as ۱۲٥۸ ۱۷۹ ۱۲۰ | a = no-frac ۸۲ ٤۳ ۷ | b = closed ۲۸ ۱٤ ۲ | c = opened	a b c <-- classified as ٥۸۷ ٤۱٥ ۱۱٥۷ | a = no-frac ۰ ۰ ۳ | b = closed ۷٤ ٤٦ ٦۷ | c = opened
Correctly Classified Percent(%)	Total	٦۱,٦٦۰٦ %	٤٦,٦۸٤۸ %	۷٥,۱۸۷٥ %	۲۷,۸٤۱٦ %
	No-frac	۸۲,٥	٤٦,۸	۸۰,۸	۲۷,۲
	closed	۱٤,۷	٥۲,۲	۳۲,٦	۰
	opened	۰	۹,٥	٤,٥	۳٥,۸

جدول ۶-۳ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۵

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱۳٤٤ ۱۱۸ ۷۰ | a = no-frac ٥٥٥ ٥٦ ۳٦ | b = closed ۲۳ ۲ ۰ | c = opened	a b c <-- classified as ۱۷٤٥ ۱۹۳۸ ۲٤۸ | a = no-frac ٤۳ ٤۷ ۰ | b = closed ۷ ۱۲ ۲ | c = opened	a b c <-- classified as ۱۲۷٥ ۱٤۲ ۱٤۰ | a = no-frac ۸۲ ٤۱ ۹ | b = closed ۲۷ ۱٤ ۳ | c = opened	a b c <-- classified as ۷۰۰ ٤۰۲ ۱۰٥۷ | a = no-frac ۰ ۰ ۳ | b = closed ۷٤ ٤۱ ۷۲ | c = opened
Correctly Classified Percent(%)	Total	٦۳,٥۲۰۹ %	٤٤,۳۸٤ %	۷٦,۱۱۰۸ %	٥۰,۳۲۷۱ %
	No-frac	۸۷,۷	٤٤,٤	۸۱,۹	۳۲,٤
	closed	۸,۷	٥۲,۲	۳۱,۱	۰
	opened	۰	۹,٥	٦,۸	۳۸,٥

با یک مقایسه‌ی کوچک میان دو جدول ۶-۳ و ۶-۲ می‌بینیم که تفاوت چندانی میان نتایج حاصل از دو سایز پنجره‌ی ۳و ۵ وجود ندارد.

جدول ۶-۴ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۷

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱٤۰۱ ۸٤ ٤۷ | a = no-frac ۵۷۱ ٤۳ ۳۳ | b = closed ۲۲ ۳ ۰ | c = opened	a b c <-- classified as ۱۷۱٤ ۱۸۸۳ ۳۳٤ | a = no-frac ٤۵ ٤۵ ۰ | b = closed ۷ ۱۲ ۲ | c = opened	a b c <-- classified as ۱۲۵۳ ۱۵٦ ۱٤۸ | a = no-frac ۷۵ ٤۵ ۱۲ | b = closed ۲۵ ۱۵ ٤ | c = opened	a b c <-- classified as ۷۸٤ ٤۰۱ ۹۷٤ | a = no-frac ۰ ۰ ۳ | b = closed ۷۷ ۳۹ ۷۱ | c = opened
Correctly Classified Percent(%)	Total	٦۵,۵۱۷۲ %	٤۳,۵٦۷۵ %	۷۵,۱۲۹۸ %	٤٤,۳۹۲۵ %
	No-frac	۹۱,٤	٤۳,٦	۸۰,۵	۳٦,۳
	closed	٦,٦	۵۰,۰	۳٤,۱	۰
	opened	۰	۹,۵	۹,۱	۳۸

جدول ۶-۵ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۹

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱٤۳۱ ۵۹ ٤۲ | a = no-frac ۵۷٦ ۳٤ ۳۷ | b = closed ۲۵ ۰ ۰ | c = opened	a b c <-- classified as ۱۷۰۲ ۱۸۵۲ ۳۷۷ | a = no-frac ٤۷ ٤۳ ۰ | b = closed ۵ ۱۳ ۳ | c = opened	a b c <-- classified as ۱۱۸۵ ۱۹٦ ۱۷٦ | a = no-frac ٦۹ ٤۹ ۱٤ | b = closed ۲۰ ۱٤ ۱۰ | c = opened	a b c <-- classified as ۸۲۷ ٤۱۵ ۹۱۷ | a = no-frac ۰ ۰ ۳ | b = closed ۸۱ ۳۵ ۷۱ | c = opened
Correctly Classified Percent(%)	Total	٦٦,٤۷۰۱ %	٤۳,۲٤۵۹ %	۷۱,۷۸۳ %	۳۸,۲۲۹ %
	No-frac	۹۳,٤	٤۳,۳	۷٦,۱	۳۸,۳
	closed	۵,۳	٤۷,۸	۳۷,۱	۰
	opened	۰	۱٤,۳	۲۲,۷	۳,۸

با مقایسه‌ی جدول ۶-۵ با جدول‌های قبلی خواهیم می‌بینیم که نتایج مربوط به چاه ۷ بهبود داشته. نتایج مربوط به چاه ۵ نیز اگرچه کمی کاهش داشته ولی با افزایش توازن همراه بوده که بسیار ارزشمند می‌باشد.

جدول ۶-۶ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۱۱

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱٤٤٥ ٥۱ ۳٦ | a = no-frac ٥۸٦ ۳۰ ۳۱ | b = closed ۲٥ ۰ ۰ | c = opened	a b c <-- classified as ۱٦۹۱ ۱۸۰۲ ٤۳۸ | a = no-frac ٤٦ ٤۳ ۱ | b = closed ٤ ۱۳ ٤ | c = opened	a b c <-- classified as ۱۱۰۹ ۲۰۸ ۲٤۰ | a = no-frac ۷۰ ٤۲ ۲۰ | b = closed ۲۰ ۱۲ ۱۲ | c = opened	a b c <-- classified as ۸۱۰ ٤۰۸ ۹٤۱ | a = no-frac ۰ ۰ ۳ | b = closed ۷٦ ۳٤ ۷۷ | c = opened
Correctly Classified Percent(%)	Total	٦٦,۹۲۳۸ %	٤۲,۹۹۸٥ %	٦۷,۱۰۹۱ %	۳۷,۷٦۰۷ %
	No-frac	۹٤,۳	٤۳	۷۱,۲	۳۷,٥
	closed	٤,٦	٤۷,۸	۳۱,۸	۰
	opened	۰	۱۹	۲۷,۳	٤۱,۲

جدول ۶-۷ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۱۳

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱٤٥٤ ٤٤ ۳٤ | a = no-frac ٥۹۰ ۲۸ ۲۹ | b = closed ۲٥ ۰ ۰ | c = opened	a b c <-- classified as ۱٦۳٦ ۱۷۸۳ ٥۱۲ | a = no-frac ٤۷ ٤۲ ۱ | b = closed ۱ ۱٦ ٤ | c = opened	a b c <-- classified as ۱۰۱٦ ۲۱۳ ۳۲۸ | a = no-frac ٥۹ ٤۱ ۳۲ | b = closed ۱۷ ۱۳ ۱٤ | c = opened	a b c <-- classified as ۸۱۲ ۳۷۹ ۹٦۸ | a = no-frac ۰ ۰ ۳ | b = closed ۷۱ ۳٥ ۸۱ | c = opened
Correctly Classified Percent(%)	Total	٦۷,۲٤۱٤ %	٤۱,٦۱۳۱ %	٦۱,۸۰۰۳ %	۳۸,۰۱٦۲ %
	No-frac	۹٤.۹	٤۱,٦	٦٥,۳	۳۷.٦
	closed	٤.۳	٤٦,۷	۳۱.۱	۰
	opened	۰	۱۹	۳۱,۸	٤۳,۳

جدول ۶-۸ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۱۵

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱٤٤۱ ٥۲ ۳۹ | a = no-frac ٥۹۸ ۲٥ ۲٤ | b = closed ۲٥ ۰ ۰ | c = opened	a b c <-- classified as ۱٥٦۹ ۱۷٥۳ ٦۰۹ | a = no-frac ٤۸ ٤۱ ۱ | b = closed ۱ ۱٦ ٤ | c = opened	a b c <-- classified as ۹۱۰ ۲۰٥ ٤٤۲ | a = no-frac ٤٥ ۳۹ ٤۸ | b = closed ۱٦ ۱۳ ۱٥ | c = opened	a b c <-- classified as ۸۱۱ ۳۸۰ ۹٦۸ | a = no-frac ۰ ۰ ۳ | b = closed ٦۹ ۳٦ ۸۲ | c = opened
Correctly Classified Percent(%)	Total	٦٦,٥۱٥٤ %	۳۹,۹۳۰۷ %	٥٥,٦۲٦۱ %	۳۸,۰۱٦۲ %
	No-frac	۹٤,۱	۳۹,۹	٥۸,٤	۳۷.٦
	closed	۳,۹	٤٥,٦	۲۹,٥	۰
	opened	۰	۱۹	۳٤,۱	٤۳,۹

جدول ۶-۹ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۱۷

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱٤٤٤ ٤۹ ۳۹ | a = no-frac ٦۰٥ ۱۷ ۲٥ | b = closed ۲٥ ۰ ۰ | c = opened	a b c <-- classified as ۱٥۲٦ ۱۷۲۰ ٦۸٥ | a = no-frac ٤۷ ٤۱ ۲ | b = closed ۱ ۱٦ ٤ | c = opened	a b c <-- classified as ۸۲۱ ۲۰٤ ٥۳۲ | a = no-frac ٤۱ ۳۹ ٥۲ | b = closed ۱٤ ۱۳ ۱۷ | c = opened	a b c <-- classified as ۸٥۳ ۳۸۲ ۹۲٤ | a = no-frac ۰ ۰ ۳ | b = closed ۷٤ ۳٦ ۷۷ | c = opened
Correctly Classified Percent(%)	Total	٦٦,۲۸۸٦ %	۳۸,۸٦٦۹ %	٥۰,٦۰٥۹ %	۳۸,۰۱٦۲ %
	No-frac	۹٤.۳	۳۸,۸	٥۲,۷	۳۹,٥
	closed	۲,٦	٤٥,٦	۲۹,٥	۰
	opened	۰	۱۹	۳۸,٦	٤۱,۲

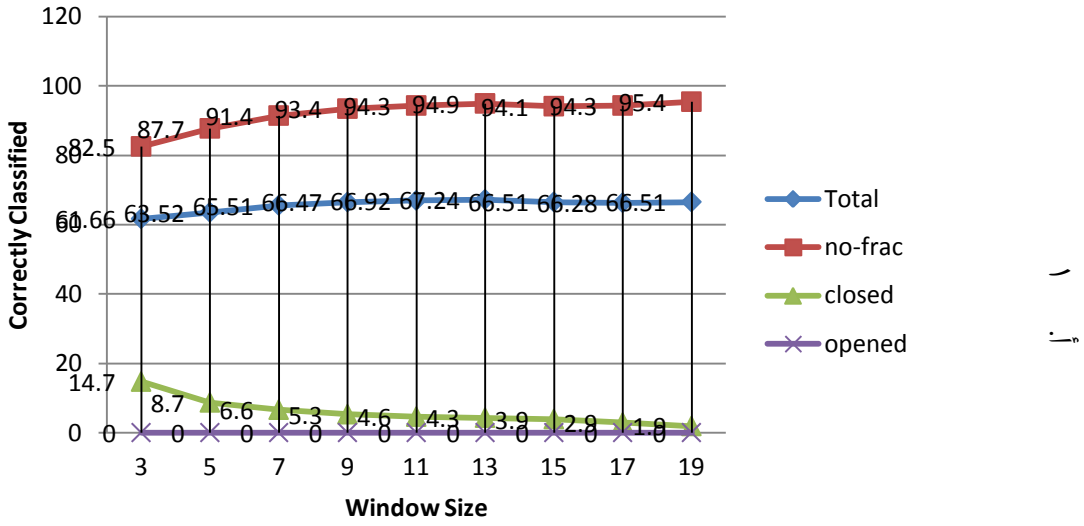
جدول ۶-۱۰ تنها مشخصه‌ی استفاده شده در الگوریتم Naïve Bayes عبارت است از ممان دوم مشخصه‌های اصلی با طول پنجره‌ی ۱۹

شماره چاه		۱	۵	۷	۸
Confusion Matrix		a b c <-- classified as ۱٤٦۲ ۳۷ ۳۳ | a = no-frac ٦۱٥ ۱۲ ۲۰ | b = closed ۲٤ ۱ ۰ | c = opened	a b c <-- classified as ۱٥۱٦ ۱٦۸۲ ۷۳۳ | a = no-frac ٤۸ ۳۹ ۳ | b = closed ۲ ۱٦ ۳ | c = opened	a b c <-- classified as ۷٥۹ ۱۹۳ ٦۰٥ | a = no-frac ٤۰ ۳٥ ٥۷ | b = closed ۱٤ ۱۲ ۱۸ | c = opened	a b c <-- classified as ۹۰٥ ٤۰۲ ۸٥۲ | a = no-frac ۰ ۰ ۳ | b = closed ۷۸ ۳۹ ۷۰ | c = opened
Correctly Classified Percent(%)	Total	٦٦,٥۱٥٤ %	۳۸,٥٤٥۳ %	٤٦,۸٥٥۲ %	٤۱,٥۰۷ %
	No-frac	۹٥,٤	۳۸,٦	٤۸,۷	٤۱,۹
	closed	۱,۹	٤۳,۳	۲٦,٥	۰
	opened	۰	۱٤,۳	٤۰,۹	۳۷,٤

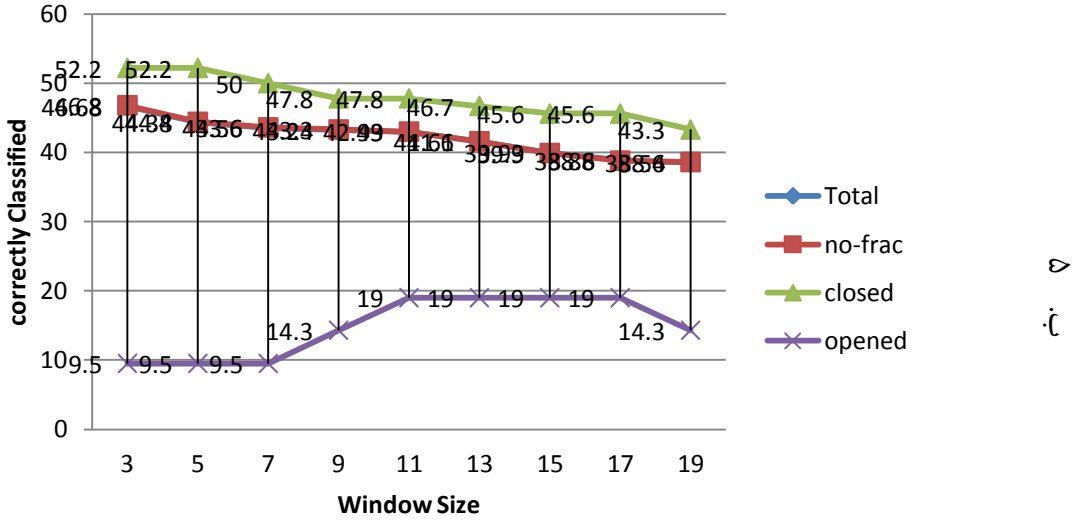
در جداول ۶-۲ تا ۶-۱۰ نتایج دسته‌بندی به صورت ماتریس درهم‌ریختگی آورده شده است و شما با دقت در این ماتریس‌ها و نیز با دقت در درصدهای صحت عملکرد می‌توانید به میزان کارائی عمل دسته‌بندی پی ببرید. هر یک از این جداول در بر دارنده‌ی اعمال الگوریتم Naïve Bayes بر روی داده‌ها می‌باشد با این تفاوت که در هر جدول طول پنجره‌ی لغزنده در مرحله‌ی پنجره‌گذاری متفاوت بوده است. بعنوان مثال در جدول ۶-۶ می‌توان مشاهده نمود که عمل پنجره‌گذاری با طول پنجره‌ی ۱۱ انجام شده و نتیجه‌ی بسیار خوبی را برای چاه ۷ به همراه داشته چرا که موازنه‌ی نسبتا خوبی میان هر سه کلاس ایجاد گردیده است.

در پایان بمنظور انجام مقایسه‌ی دقیق و یافتن مناسب‌ترین طول پنجره، خلاصه نتایج جداول ۶-۲ تا ۶-۱۰ در نمودارهای جدول ۶-۱۱ به صورت گرافیکی ترسیم شده‌اند.

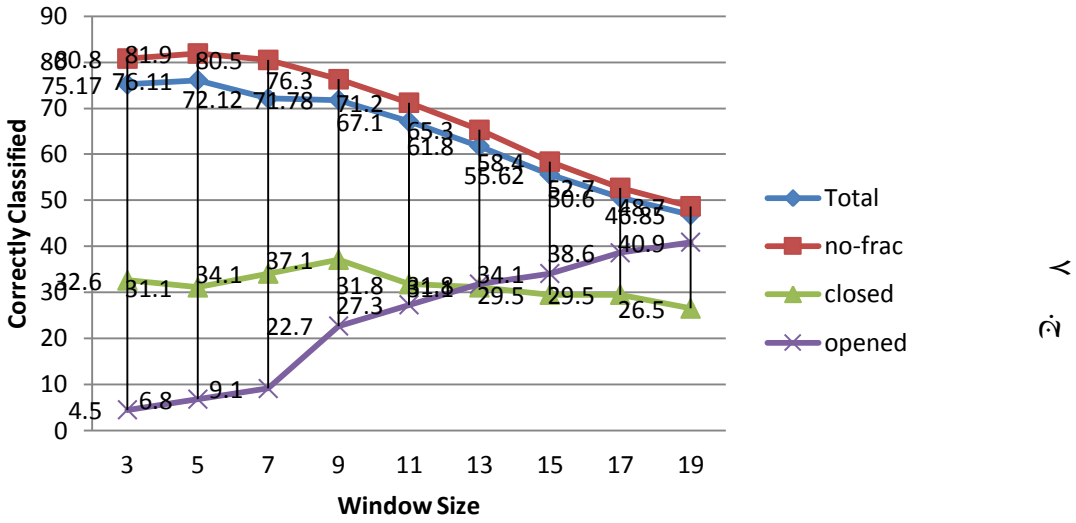
جدول ۶-۱۱ مقایسه میزان کارائی طول‌های مختلف پنجره‌گذاری واریانس در الگوریتم Naïve Bayes



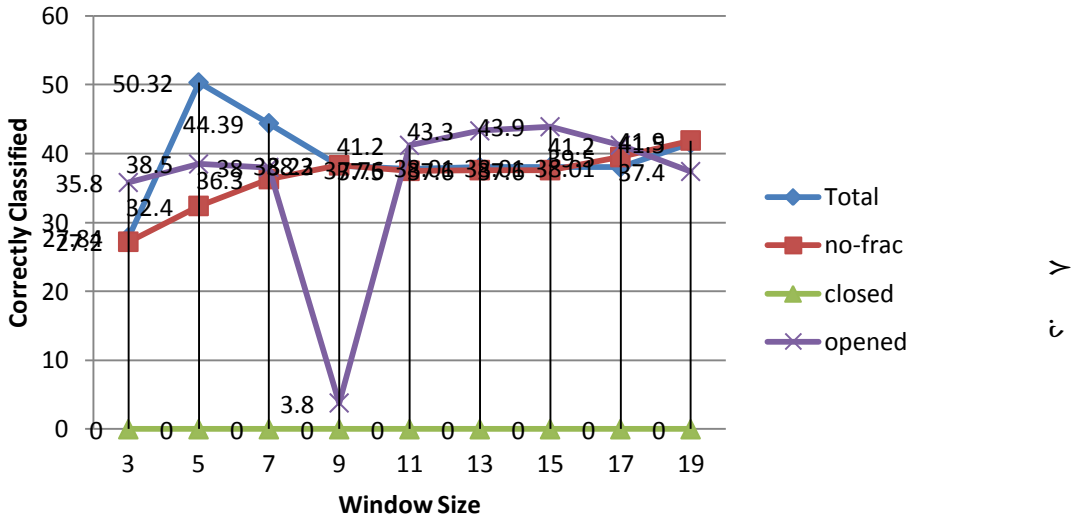
ب-



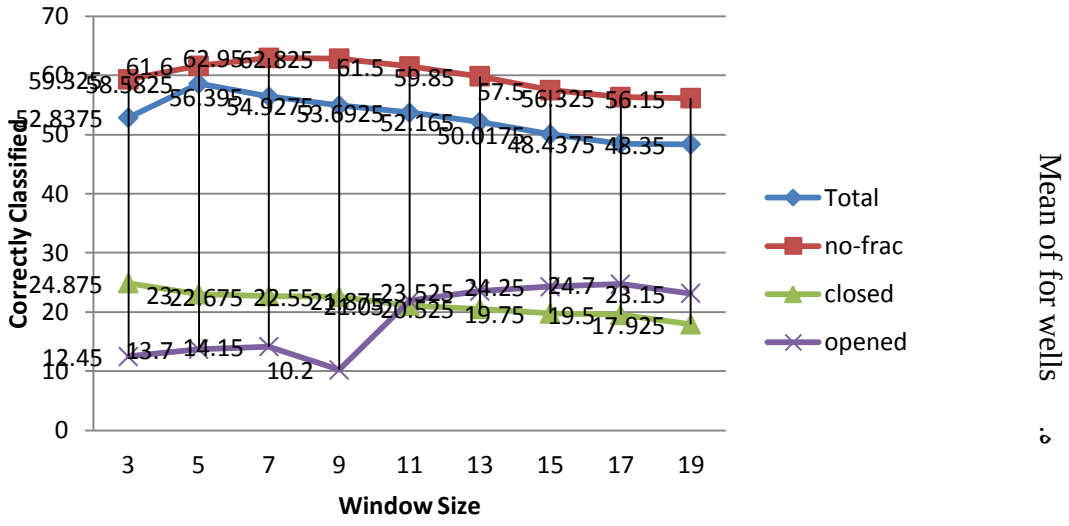
ج-


Correctly Classified
Window Size
Total
no-frac
closed
opened

شكل ٢: ٧


Correctly Classified
Window Size
Total
no-frac
closed
opened

شكل ٣: ٧



همانطور که در نمودارهای جدول ۶-۱۱ مشاهده می‌شود یافتن مناسب‌ترین طول پنجره برای الگوریتم Naïve Bayes کار ساده‌ای نیست، بعنوان مثال در قسمت a که مربوط به چاه ۱ می‌باشد می‌توان طول پنجره‌ی ۳ را بعنوان مناسب‌ترین طول در نظر گرفت، در قسمت b که مربوط به چاه ۵ می‌باشد طول پنجره‌ی ۱۱ عملکرد بهتری را از خود نشان داده است، ولی در چاه‌های ۷ و ۸ که نمودارهای آن‌ها به ترتیب در قسمت‌های cو d آمده است نمی‌توان یک و یا چند طول پنجره‌ی خاص را بعنوان مناسب‌ترین طول در نظر گرفت بعنوان مثال در چاه ۷ می‌توان عدد ۱۷ و یا ۹ را بعنوان مناسب‌ترین طول انتخاب نمود که البته بستگی به این دارد که تا چه حد بخواهیم کارائی کل را فدای توازن میان کلاس‌ها گردانیم. در قسمت e که میانگینی از چهار نمودار قبلی می‌باشد نیز می‌توانیم طول پنجره‌ی ۱۱ را بعنوان مناسب‌ترین طول پنجره انتخاب نمائیم.

ج. انتخاب محل مناسب در هر پنجره برای جایگزین شدن با واریانس

برای انتخاب داده‌ی جایگزین شونده با واریانس در هر پنجره می‌توان یکی از سه روش زیر را انتخاب نمود که این انتخاب کاملا وابسته به ماهیت داده‌ها بوده و در هر مسأله باید جداگانه مورد آزمایش قرار گیرد و لذا ما نیز در اینجا این آزمایش را انجام داده‌ایم. این سه روش عبارتند از:

۱. انتخاب داده‌ی اول در هر پنجره (نگاه به آینده)
۲. انتخاب داده‌ی وسط در هر پنجره (نگاه به آینده و گذشته)
۳. انتخاب داده‌ی آخر در هر پنجره (نگاه به گذشته)

در آزمایشات صورت گرفته مشخص گردید که روش دوم بهترین نتیجه را در بر داشته. و لذا در قسمت ۴-۲-۲-۵ نتایج دسته بندی بر این اساس گزارش شده است.

۶-۳-۴- نرمال‌سازی داده‌ها

در دو آزمایش مجزا یکبار قبل از اجرای دسته‌بندی بر روی داده‌های بدست آمده از واریانس پنجره-گذاری، نرمالسازی انجام شد و بار دیگر نرمالسازی انجام نشد. البته نکته‌ی حائز اهمیت که در امر نرمالسازی عمومیت داشته و ما نیز در اینجا ملزم به در نظرگرفتن آن هستیم اینست که برای نرمالسازی باید همه‌ی داده‌ها با هم و بعنوان یک سیگنال واحد نرمالسازی شوند.

نتیجه‌ی بدست آمده از این آزمایش کاملا با ماهیت الگوریتم Naïve Bayes و Random Tree مطابقت دارد و آن اینکه این الگوریتم نسبت به نرمال بودن و یا نبودن داده‌ها حساسیت نداشته و در آخر نتیجه‌ی هر دو آزمایش کاملا یکسان بوده است که ما بمنظور استانداردسازی بیشتر از این پس در آزمایشاتمان از سیگنال نرمال‌سازی شده استفاده می‌نمائیم.

۶-۳-۵- استفاده از عمق بعنوان یک مشخصه

در ادامه، عمق نیز به عنوان یک ویژگی مد نظر قرار گرفت چرا که از نقطه نظر زمین شناسی با توجه به این نکته که چاه‌های مورد آزمایش از لحاظ موقعیت جغرافیایی به هم نزدیک بوده و در یک میدان نفتی قرار گرفته‌اند بنابراین ممکن است هر یک از سطوح مختلف پوسته‌ی زمین در تمامی این چاه‌ها دارای شباهتهایی باشد. البته با توجه به نتایج بدست آمده از این آزمایش مشخص شد که این رویکرد نه تنها منجر به بهبود عملکرد نشد بلکه باعث کاهش کارائی نیز گردید.

۶-۳-۶- استفاده از داده‌های اولیه به همراه داده‌های حاصل از پنجره-گذاری

در این مرحله از این تئوری ایده گرفتیم که تعداد بیشتر مشخصه‌ها می‌تواند به فرآیند دسته‌بندی کمک نماید. بنابراین این بار آزماشات قبل را با استفاده از ۲۲ مشخصه، شامل داده‌های خام و پردازش‌شده انجام دادیم. که البته در پایان آزمایشات مشخص شد که با توجه به همبستگی زیاد میان داده‌های خام و داده‌های پردازش‌شده این کار به گمراهی الگوریتم دسته‌بندی و در نتیجه کاهش کارائی فرآیند دسته بندی منجر گردیده است. بعلت خارج بودن از حد تحمل این مستند، از ارائه‌ی نتایج این آزمایش در اینجا صرف نظر می‌کنیم.

۶-۴- انتخاب حداکثری تعداد نمونه‌های آزمایش

در این گام از انجام پروژه تاکید بر افزایش تعداد نمونه‌ها می‌باشد و این بدان معناست که باید تعداد چاه‌های بیشتری را در نظر بگیریم، این موضوع ما را بر آن می‌دارد که برخی مشخصه‌هایی که میان چاه‌های مختلف مشترک نیستند را نادیده گرفته و در دسته‌بندی از آن‌ها کمک نگیریم.

۶-۴-۱- گزینش مشخصه‌ها و داده‌ها

در جدول ۶-۱۲ مشخص است که هفت چاه‌نمودار Caliper, Gamma ray (SGR, CGR), sonic(DT), density (PEF , RHOB), neutron (NPHI)، در میان هفت چاه حضور دارند و لذا ما از این چاه‌نمودارها بعنوان مشخصه‌های داده و از داده‌های متعلق به هفت چاه به شماره‌های ۱، ۲، ۳، ۵، ۶، ۷ و ۸ بعنوان نمونه‌های آزمایش بهره می‌گیریم.

البته همانطور که در جدول مشخص است می‌توان از مشخصه‌های SGRو NPHI نیز صرف نظر نموده و داده‌های چاه شماره‌ی ۴ را نیز به خدمت گرفت که البته با استفاده از یک الگوریتم ارزیابی میزان کارائی مشخصه‌ها مشخص گردید این دو مشخصه از اهمیت زیادی برخوردار هستند و لذا ترجیح بر آن شد که این دو مشخصه حذف نشوند هر چند داده‌های یک چاه را از دست خواهیم داد. پس به دو نحو می‌توان داده‌ها را انتخاب نمود:

۱. انتخاب هفت چاه که هر یک دارای هفت چاه‌نمودار می‌باشند و این هفت چاه‌نمودار عبارتنداز:
CALI, DT, CGR, SGR, RHOB, NPHI, PEF
۲. انتخاب هشت چاه که هر کدام دارای پنج چاه نمودار می‌باشند که عبارتند از همان ترکیب بالا به غیر از NPHI و SGR

جدول ۶-۱۲ چاه‌نمودارهای موجود در هر چاه

چاه \ ویژگی	CALI	DT	CGR	SGR	RHOB	NPHI	RT	FLAG	PEF	GR	POTA	URAN	THOR	PHIE	SWE	CALCITE	DOLOMITE	ANHYDRITE	SHALE	ILM	ILD	MSFL	LLD	LLS
۱	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
۲	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
۳	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
۴	✓	✓	✓	✗	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
۵	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗
۶	✓	✓	✓	✓	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗
۷	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✗	✗	✗
۸	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓	✓	✓	✗	✗	✗

همانطور که گفته شد محاسباتی لازم است تا بتوانیم از میان دو انتخاب بالا بهترین را گزینش نمائیم. بدین منظور از الگوریتم InfoGainAttributeEval استفاده نمودیم که در فصل دوم و در قسمت ۲-۱ توضیح داده شده است. این الگوریتم را بر روی داده‌های ادغام شده چاه‌هایی که نمودارهای NPHI و SGR آنها در دست هست اعمال نمودیم. که نتایج آن را در جدول ۶-۱۳ ملاحظه می‌فرمائید.

با مشاهده‌ی جدول ۶-۱۳ در می‌یابیم که NPHI از اهمیت فوق‌العاده‌ای در کلاس‌بندی برخوردار است و به این سادگی نمی‌توان از آن گذشت. پس نتیجه می‌گیریم به خاطر آنکه تعداد نمونه‌ها را افزایش دهیم یا به عبارت واضح‌تر به خاطر افزایش تعداد چاه‌ها را از هفت حلقه به هشت حلقه نباید

از اهمیت NPHI بگذریم، پس در نتیجه تصمیم می‌گیریم از هفت حلقه چاه که دارای هفت صفت مشترک هستند استفاده کنیم.

جدول ۶-۱۳ رتبه‌بندی نمودارها از نظر تاثیرگذاری بر روی کلاس‌بندی

	Attribute	Rank
۱	NPHI	۰,۰۲۸۵۲
۲	CALI	۰,۰۲۴۷
۳	RHOB	۰,۰۲۳۷۵
۴	PEF	۰,۰۱۹۵۹
۵	DT	۰,۰۰۹۹۴
۶	CGR	۰,۰۰۵۶۴
۷	SGR	۰,۰۰۳۵۷

۶-۴-۲- استفاده از داده‌های خام

هر چند استفاده از داده‌های خام نتیجه‌ی قابل قبولی را در پی نداشت ولی تا حد زیادی امیدوار کننده بود چرا که با همان الگوریتم به کار رفته در قسمت ۴-۲ یعنی Naïve Bayes کارائی دسته‌بندی به این شکل بود که اغلب داده‌های متعلق به کلاس no-frac را به درستی تشخیص داده و داده‌های کلاس closed را نیز در نیمی از چاه‌ها حدود ۱۴ درصد به درستی تشخیص می‌دهد و این نتیجه ما را برای اجرای گام‌های بعدی امیدوارتر می‌نماید.

۶-۴-۳- انتخاب الگوریتم دسته‌بندی

با توجه به اینکه ماهیت داده‌ها تغییر نکرده و نیز با توجه به اینکه حجم داده‌ها از نظر آماری هنوز هم نمونه‌ی کاملی از جامعه‌ی هدف (داده‌های پتروفیزیکی چاه) نمی‌باشد، بنابراین همانند قسمت ۴-۲-۲ در اینجا نیز از الگوریتم‌های Random Treeو Naïve Bayes استفاده می‌کنیم.

۶-۴-۴- ایجاد مشخصه‌های جدید با استفاده از تکنیک پنجره‌گذاری

همانطور که در قسمت ۴-۳-۲ مطرح گردید نتایج دسته بندی با داده‌های خام هر چند امیدوارکننده بود ولی کاملا مشخص است که داده‌ها هنوز آماده‌ی اجرای یک دسته بندی موفق نمی‌باشند، بنابراین ما باید به دنبال انجام یکی از روشهای مطرح در پیش‌پردازش باشیم که همانا استخراج مشخصه‌های جدید از روی مشخصه‌های داده‌های خام می‌باشد. در این قسمت قصد داریم به جای داده‌های خام از داده‌های حاصل از پنجره‌گذاری استفاده نمائیم. (برای توضیحات بیشتر می‌توانید به بخش ۴-۳-۳ مراجعه نمائید). در ادامه‌ی این قسمت از کلیه‌ی تجربیات بدست آمده از قسمت ۶-۲ استفاده شده و سپس با ارائه‌ی نتایج آزمایشات قسمت ۶-۲ و نیز ۶-۳ به بررسی و مقایسه‌ی نتایج می‌پردازیم.

۶-۵- نتایج دسته‌بندی به سه کلاس

همانطور که تا اینجا ملاحظه فرمودید می‌توان با دو رویکرد متفاوت به انتخاب داده‌ها و مشخصه‌ها پرداخت که در نتیجه‌ی آن دو دسته پاسخ خواهیم داشت. این دو رویکرد به طور خلاصه عبارت بودند از یکی انتخاب داده‌های ۷ چاه با ۷ مشخصه و دیگری انتخاب داده‌های ۴ چاه با ۱۱ مشخصه.

در قسمت‌های بعدی خواهیم دید که رویکرد دوم یعنی استفاده از ۴ چاه با هفت ویژگی نتایج بهتری را در دسته بندی سه‌کلاسه نتایج بهتری را به دنبال خواهد داشت. ضمن این‌که کارامدترین الگوریتم نیز Majority Voting میان Naïve Bayes و Random Tree با K=۱۰ شناسایی گردید.

۶-۵-۱- نتایج استفاده از داده‌های ۴ چاه با ۱۱ مشخصه در دسته‌بندی سه کلاسه

جدول ۶-۱۴ بهترین دسته بندی سه کلاسه با استفاده از داده‌های ۴ چاه و ۱۱ مشخصه، به صورت بهینه سازی مجزا برای هر چاه

classifier	parameteres	Confusion Matrix	Correctly Classified Percent(%)				well
			Total	no-frac	closed	opened	
Random Tree	K=۱۰	a b c <-- classified as ۱۲۱٤ ۳۸ ۲۸۰ | a = no-frac ٤۷٥ ٤۱ ۱۳۱ | b = closed ۱۷ ۳ ٥ | c = opened	٥۷,۱٦۸۸ %	۷۹,۲	٦,۳	۲۰	۱
Majority Voting (NB+RT۱۰)		a b c <-- classified as ۲٤۲۲ ۹۹۰ ٥۱۹ | a = no-frac ٥۱ ۳۰ ۹ | b = closed ۹ ٦ ٦ | c = opened	٦۰,۸۱۱٥ %	٦۱,٦	۳۳,۳	۲۸,٦	٥
Majority Voting (RT۳+NB)		a b c <-- classified as ۱۲۹۳ ۸٤ ۱۸۰ | a = no-frac ۹۷ ۲۲ ۱۳ | b = closed ۳۱ ۸ ٥ | c = opened	۷٦,۱٦۸٥ %	۸۳,۰	۱٦,۷	۱۱,٤	۷
Majority Voting (RT۱۰+NB)		a b c <-- classified as ۱۳۱٥ ۳۲٦ ٥۱۸ | a = no-frac ۰ ۲ ۱ | b = closed ۱۲٤ ۳٥ ۲۸ | c = opened	٥۷,۲٥۸٤ %	٦۰,۹	٦٦,۷	۱,٥	۸

جدول ۶-۱۵ بهترین دسته بندی سه کلاسه با استفاده از داده‌های ۴ چاه و ۱۱ مشخصه، به صورت مشترک و قابل تعمیم

classifier	parameteres	Confusion Matrix	Correctly Classified Percent(%)				well
			Total	no-frac	closed	opened	
Majority Voting (NB+RT۱۰)	K=۱۰	a b c <-- classified as ۱۳۲۷ ۲۵ ۱۸۰ | a = no-frac ۵۲۶ ۲۵ ۹۶ | b = closed ۱۹ ۳ ۳ | c = opened	۶۱,۴۷۹۱ %	۸۶,۶	۳,۹	۱۲,۰	۱
Majority Voting (NB+RT۱۰)	K=۱۰	a b c <-- classified as ۲۴۲۲ ۹۹۰ ۵۱۹ | a = no-frac ۵۱ ۳۰ ۹ | b = closed ۹ ۶ ۶ | c = opened	۶۰,۸۱۱۵ %	۶۱,۶	۳۳,۳	۲۸,۶	۵
Majority Voting (NB+RT۱۰)	K=۱۰	a b c <-- classified as ۱۳۲۲ ۸۳ ۱۵۲ | a = no-frac ۹۵ ۲۳ ۱۴ | b = closed ۳۱ ۹ ۴ | c = opened	۷۷,۸۴۱۹ %	۸۴,۹	۱۷,۴	۹,۱	۷
Majority Voting (NB+RT۱۰)	K=۱۰	a b c <-- classified as ۱۳۱۵ ۳۲۶ ۵۱۸ | a = no-frac ۰ ۲ ۱ | b = closed ۱۲۴ ۳۵ ۲۸ | c = opened	۵۷,۲۵۸۴ %	۶۰,۹	۶۶,۷	۱,۵	۸

۶-۵-۲- نتایج استفاده از داده‌های ۷ چاه با ۷ مشخصه در دسته‌بندی سه کلاسه

در جدول‌های این قسمت نتایج دسته‌بندی داده‌ها در سه کلاس با استفاده از ۷ چاه‌نمودار آورده شده است.

جدول ۶-۱۶ بهترین دسته بندی سه کلاسه با استفاده از داده‌های ۷ چاه و ۷ مشخصه، به صورت بهینه سازی مجزا برای هر چاه جدول

classifier	parameteres	Confusion Matrix	Correctly Classified Percent(%)				well
			Total	no-frac	closed	Opened	
Random Tree	K=۵	a b c <-- classified as ۱۳۱۳ ۱۳٤ ۸۵ | a = no-frac ۵۱٦ ۹٦ ۳۵ | b = closed ۱۵ ۷ ۳ | c = opened	٦۲,۷۹٤۹ %	۸۵,۷	۱٤,۸	۱۲	۱
Random Tree	K=۷ or ۹	a b c <-- classified as ۲۳٤۸ ۱۹۵ ۱٤۵ | a = no-fra ٦۰۳ ٦۷ ۲۱ | b = closed ۹۲ ۹ ٤ | c = opened	٦۹,٤۳۱۷ %	۸۷,٤	۹,۷	۳,۸	۲
Majority Voting (RT۵+NB)	K=۵	a b c <-- classified as ۲۸۳۱ ۱٤۲۲ ۹۰ | a = no-fra ۵۵ ۱۸ ۰ | b = closed ٤۱ ۱۱ ۱ | c = opened	٦۳,۷۷۲۷ %	٦۵,۲	۲٤,۷	۱,۹	۳
Majority Voting (RT۳+NB)	K=۳	a b c <-- classified as ۲۸۷۹ ۹٤۱ ۱۱۱ | a = no-fra ۵۷ ۲۷ ٦ | b = closed ۱۳ ٦ ۲ | c = opened	۷۱,۹٤٤٦ %	۷۳,۲	۳۰,۰	۹,۵	۵
Majority Voting (RT۱+NB)	K=۱	a b c <-- classified as ۱۵۹۹ ۸۹۳ ۵۰ | a = no-frac ۱٦٤ ۱۲۲ ۹ | b = closed ۸۵ ٦۱ ۵ | c = opened	۵۷,۷٦٤٤ %	٦۲,۹	٤۱,٤	۳,۳	٦
Random Tree	K=۷ or ۹	a b c <-- classified as ۱۳٦٦ ۱۰٦ ۸۵ | a = no-frac ۱۲۱ ۸ ۳ | b = closed ۳۹ ۳ ۲ | c = opened	۷۹,۳۹۹۹ %	۸۷,۷	٦,۱	٤,۵	۷
Random Tree	K=۰	a b c <-- classified as ۱۹۱٦ ۱۷۲ ۷۱ | a = no-frac ۱ ۲ ۰ | b = closed ۱۷۳ ۱۱ ۳ | c = opened	۸۱,۷۷۹۵ %	۸۸,۷	٦٦,۷	۱,٦	۸

جدول ۶-۱۷ بهینه سازی پارامترها در دسته‌بندی سه کلاسه با استفاده از داده‌های ۷چاه و ۷ مشخصه، برای همه ی چاه‌ها به صورت مشترک و قابل تعمیم

classifier	parameteres	Confusion Matrix	Correctly Classified Percent(%)				Well
			Total	no-frac	closed	Opened	
Majority Voting(RT۳+NB)	K=۳	a b c <-- classified as ۱٤۰٦ ۷۷ ٤۹ | a = no-frac ٥۷٥ ٤۷ ۲٥ | b = closed ۲۰ ٤ ۱ | c = opened	٦٥,۹۷۱ %	۹۱,۸	۷,۳	٤	۱
Majority Voting(RT۳+NB)	K=۳	a b <-- classified as ۲٥۲۲ ۱٦٦ | a = no-frac ۷٤٥ ٥۱ | b = fracture	۷۲,٦٤٦ %	۹۲,۷	٥,۲	۳,۸	۲
Majority Voting(RT۳+Naive)	K=۳	a b c <-- classified as ۲۸٤۰ ۱٤۰۷ ۹٦ | a = no-frac ٥۷ ۱٤ ۲ | b = closed ٤۰ ۱۲ ۱ | c = opened	٦۳,۸۸٤ %	٦٥,٤	۱۹,۲	۱,۹	۳
Majority Voting(RT۳+NB)	K=۳	a b <-- classified as ۲۷٤۷ ۱۱۸٤ | a = no-frac ٦۳ ٤۸ | b = fracture	۷۱,۹٤٥ %	۷۳,۲	۳۰,۰	۹,٥	٥
Majority Voting(RT۳+NB)	K=۳	a b c <-- classified as ۱٦۱۲ ۸۸۸ ٤۲ | a = no-frac ۱۷۰ ۱۱۹ ٦ | b = closed ۸۸ ٦۲ ۱ | c = opened	٥۷,۹٦٥ %	٦۳,٤	٤۰,۳	۰,۷	٦
Majority Voting(RT۳+NB)	K=۳	a b c <-- classified as ۱۳۱۷ ۲۰٤ ۳٦ | a = no-frac ۱۰۱ ۲۹ ۲ | b = closed ۳۰ ۱۲ ۲ | c = opened	۷۷,۷۸٤ %	۸٤,٦	۲۲,۰	٤,٥	۷
Majority Voting(RT۳+Naive)	K=۳	a b c <-- classified as ۱۸۹۳ ۲۲۸ ۳۸ | a = no-frac ۱ ۲ ۰ | b = closed ۱۷٦ ۹ ۲ | c = opened	۸۰,۷٥۸ %	۸۸,۷	٦٦,۷	۱,۱	۸

۶-۶- نتایج دسته‌بندی به دو کلاس

در پایان به این نتیجه رسیدیم که دسته بندی با استفاده از دو الگوریتم دسته‌بند به نام های Random Tree و Naïve Bayes و ترکیب آنها با روش Majority Voting بهترین پاسخ را به همراه دارد. به این صورت که همانند قسمت ۶-۴، دو الگوریتم دسته‌بند در رای گیری شرکت می‌کنند. البته

این که پارامتر k در Random Tree باید چه مقداری داشته باشد نیز یک سؤال مهم می‌باشد. با توجه به اینکه هنوز به یک الگوریتم مناسب برای سنجش میزان دقت عملیات دسته‌بندی دست نیافته ایم، کار سنجش میزان کارائی را به صورت دستی انجام می‌دهیم، به این صورت که در این سنجش باید همواره دو موضوع را مد نظر قرار دهیم، یکی اینکه کل داده‌ها با درصد نسبتا بالایی مثلا ۷۰ درصد به درستی دسته بندی شوند و دیگری اینکه موازنه‌ی خوبی میان میزان دسته بندی صحیح میان همه‌ی دسته‌ها وجود داشته باشد.

بنابراین بهترین الگوریتم‌های دسته بندی و نیز بهترین ضرائب انتخاب شدند که از آن میان برای انتخاب بهترین ضریب k از میان سه عدد ۰و ۶ و ۷ ناگذیر به اجرای میانگین‌گیری هستیم. که با توجه به جدول ۶-۱۸، k=۶ بهترین موازنه را در بر داشته چرا که در صد بیشتری از کلاس fracture را تشخیص داده است.

نکته‌ی بسیار مهمی که در این قسمت به آن دست یافتیم اینست که در دسته‌بندی به دو کلاس برخلاف دسته‌بندی به سه کلاس، استفاده از داده‌های ۷ چاه با ۷ مشخصه نتایج بهتری را بدنبال دارد. این در حالیست که در قسمت ۶-۴ در مورد دسته‌بندی به سه کلاس استفاده از داده‌های ۴ چاه با ۱۱ مشخصه نتایج بهتری را در بر داشت.

جدول ۶-۱۸ میانگین درصد صحت دسته‌بندی دو کلاسه با استفاده از الگوریتم Random Tree با ضرایب مختلف k برای انتخاب کارامدترین ضریب k

K	Total	No-frac	fracture
۰	۶۹,۲۸۰۹	۷۶,۱۸۵۷۱	۲۵,۲۲۸۵۷
٦	۶۸,۶۸۵۲	۷۶,۵۷۱۴۳	۲۶,۱
۷	۶۹,۴۹۶۶	۷۸,۰۵۷۱۴	۲۴,۹۲۸۵۷

از آنجا که کارامدتر بودن به کارگیری داده‌های ۷ چاه کاملاً مشهود است، بنابراین نتایج استفاده از داده‌های ۴ چاه در اینجا آورده نشده است.

جدول ۶-۱۹ بهترین دسته بندی دوکلاسه به صورت بهینه سازی مجزا برای هر چاه

classifier	parameteres	Confusion Matrix	Correctly Classified Percent(%)			well
			Total	no-frac	fracture	
Random Tree	K=٦	a b <-- classified as ١٢٧٩ ٢٥٣ | a = no-frac ٥٢٥ ١٤٧ | b = fracture	٦٤,٧٠٠ %	٨٣,٥	٢١,٩	١
Random Tree	K=٧	a b <-- classified as ٢٤١٣ ٢٧٥ | a = no-frac ٧٠٢ ٩٤ | b = fracture	٧١,٩٥٧ %	٨٩,٨	١١,٨	٢
Majority Voting(RT٧+Naive)	K=٧	a b <-- classified as ٢٧٥٠ ١٥٩٣ | a = no-frac ٨٨ ٣٨ | b = fracture	٦٢,٣٨٥ %	٦٣,٣	٣٠,٢	٣
Majority Voting(NB+RT٦)	K=٦	a b <-- classified as ٢٧٤٧ ١١٨٤ | a = no-frac ٦٣ ٤٨ | b = fracture	٦٩,١٤٩ %	٦٩,٩	٤٣,٢	٥
Majority Voting(NB+RT٧)	K=٧	a b <-- classified as ١٧١٣ ٨٢٩ | a = no-frac ٢٧١ ١٧٥ | b = fracture	٦٣,١٨٦ %	٦٧,٤	٣٩,٢	٦
Naïve Bayes		a b <-- classified as ١٢٦٨ ٢٨٩ | a = no-frac ١٠٣ ٧٣ | b = fracture	٧٧,٣٨٠ %	٨١,٤	٤١,٥	٧
Majority Voting(NB+RT٧)	K=٧	a b <-- classified as ١٦٠٢ ٥٥٧ | a = no-frac ١٥٣ ٣٧ | b = fracture	٦٩,٧٧٤ %	٧٤,٢	١٩,٥	٨

در پایان در جدول ۶-۲۱ خلاصه‌ی نتایج دسته‌بندی دو کلاسه و سه کلاسه، میانگین نتایج در هفت چاه مورد آزمایش به صورت اختصاصی هر چاه و قابل تعمیم آورده شده است.

جدول ۶-۲۰ بهینه سازی پارامترها برای دسته‌بندی دو کلاسه همه ی چاه‌ها به صورت مشترک و قابل تعمیم

classifier	parameteres	Confusion Matrix	Correctly Classified Percent(%)			Well
			Total	no-frac	fracture	
Majority Voting(RT٦+NB)	K=٦	a b <-- classified as ١٢٨٠ ٢٥٢ | a = no-frac ٥٢٩ ١٤٣ | b = fracture	٦٤,٥٦٤ %	٨٣,٦	٢١,٣	١
Majority Voting(RT٦+NB)	K=٦	a b <-- classified as ٢٥٢٢ ١٦٦ | a = no-frac ٧٤٥ ٥١ | b = fracture	٧٣,٨٥٢ %	٩٣,٨	٦,٤	٢
Majority Voting(RT٦+Naive)	K=٦	a b <-- classified as ٢٧٨٤ ١٥٥٩ | a = no-frac ٩١ ٣٥ | b = fracture	٦٣,٠٧٩ %	٦٤,١	٢٧,٨	٣
Majority Voting(NB+RT٦)	K=٦	a b <-- classified as ٢٧٤٧ ١١٨٤ | a = no-frac ٦٣ ٤٨ | b = fracture	٦٩,١٤٩ %	٦٩,٩	٤٣,٢	٥
Majority Voting(NB+RT٦)	K=٦	a b <-- classified as ١٧٢٦ ٨١٦ | a = no-frac ٢٧٤ ١٧٢ | b = fracture	٦٣,٥٢٠ %	٦٧,٩	٣٧,٦	٦
Majority Voting(NB+RT٦)	K=٦	a b <-- classified as ١٢٩٢ ٢٦٥ | a = no-frac ١٢٥ ٥١ | b = fracture	٧٧,٤٩٦ %	٨٣	٢٩	٧
Majority Voting(NB+RT٦)	K=٦	a b <-- classified as ١٥٩١ ٥٦٨ | a = no-frac ١٥٧ ٣٣ | b = fracture	٦٩,١٣٦ %	٧٣,٧	١٧,٤	٨

جدول ۶-۲۱ خلاصه‌ی نتایج دسته بندی دو کلاسه سه کلاسه و میانگین نتایج

نوع دسته بندی	۲کلاسه- تعمیم پذیر	۲کلاسه- اختصاصی	۳جلاسه- تعمیم پذیر	۳کلاسه- اختصاصی
١	٦٤,٥٦٪	٦٤,٧٠٪	٦٥,٩٧٪	٦٢,٧٩٪
٢	٧٣,٨٥٪	٧١,٩٦٪	٧٢,٦٥٪	٦٩,٤٣٪
٣	٦٣,٠٨٪	٦٢,٣٩٪	٦٣,٨٨٪	٦٣,٧٧٪
٥	٦٩,١٥٪	٦٩,١٥٪	٧١,٩٥٪	٧١,٩٤٪
٦	٦٣,٥٢٪	٦٣,١٩٪	٥٧,٩٧٪	٥٧,٧٦٪
٧	٧٧,٥٠٪	٧٧,٣٨٪	٧٧,٧٨٪	٧٩,٤٠٪
٨	٦٩,١٤٪	٦٩,٧٧٪	٨٠,٧٦٪	٨١,٧٨٪
میانگین	٦٨,٦٩	٦٨,٣٦١٦	٧٠,١٣٦١	٦٩,٥٥٥٤

فصل هفتم

نتیجه گیری و پیشنهادات

۷-۱- نتیجه‌گیری

در این تحقیق توانستیم روشی برای مشخص کردن کلاس هر یک از نواحی یک چاه نفت با استفاده از داده‌های پتروفیزیکی آن چاه ارائه دهیم. هر ناحیه می‌تواند یکی از سه کلاس شکستگی باز، شکستگی بسته و شکسته‌نشده را دارا باشد. این روش بر روی داده‌های هفت چاه از چاه‌های جنوب کشور اعمال شد، که دربردارنده ۱۱ چاه‌نمودار پتروفیزیکی و نیز تفسیر چاه نمودارهای تصویری بمنظور تعیین کلاس واقعی هر ناحیه می‌باشد. نتایج بدست آمده در فصل ششم حاکی از توان بالای این روش در شناسایی نواحی شکسته می‌باشد که میانگین دقت آن بیش از ۷۰,۱ درصد را در دسته-بندی سه کلاسه‌ی قابل تعمیم و نیز بالاتر از ۶۸,۶ درصد را در دسته‌بندی دو کلاسه‌ی قابل تعمیم از خود نشان می‌دهد.

این در حالیست که تمرکز اصلی این پروژه از ابتدا بر روی دسته‌بندی سه کلاسه بوده است و دسته-بندی دو کلاسه در پایان مطالعات و صرفاً برای آندسته از خوانندگانی انجام شده که قصد مقایسه‌ی این روش با سایر روش‌های گذشته را دارند.

در این تحقیق شاهد انجام یک پروژه‌ی داده‌کاوی بودیم که از لحاظ تکنیکی دربردارنده‌ی پیچیدگی-هایی می‌باشد. یکی از پیچیدگی‌های مسأله پیچیدگی رفتاری داده‌های پتروفیزیکی و بالاخص پیچیدگی رفتاری شکستگی‌ها می‌باشد که تشخیص و بررسی آن را برای محققان و کارشناسان صنعت نفت به امری دشوار مبدل نموده است. از دیدگاه داده‌کاوی نیز این مسأله که اختلاف بسیار زیادی میان حجم داده‌های کلاس‌های مختلف وجود دارد خود به تنهائی منجر به پیچیدگی عمل دسته‌بندی می‌گردد.

۷-۲- پیشنهادات و کارهای آینده

هر چند نتایج بدست آمده در این تحقیق بسیار ارزنده و قابل ارائه در محافل علمی می‌باشد ولی برای آنکه این نتایج قابل استفاده در صنعت باشند لازم است نتایج بهتری حاصل شود. البته نباید از این نکته غافل شد که دراختیار قرار گرفتن داده‌های بیشتر مسلماً به طرز چشم‌گیری به بهبود عملکرد روش پیشنهادی کمک خواهد نمود.

از آنجا که برای ارائه‌ی روش پیشنهادی در این پروژه تعداد زیادی از الگوریتم‌های دسته‌بندی و پیش-پردازش مورد تحلیل و بررسی قرار گرفته تا از میان آنها مؤثرترین و کاربردی‌ترین روش‌ها انتخاب شوند، لذا برای ادامه‌ی این پروژه در آینده پیشنهاد می‌شود روش پیشنهادی به خوبی مورد مطالعه قرار گرفته و با توجه به زیاد بودن پارامترهای مورد استفاده در این روش، از قبیل طول پنجره، تابع هسته‌ی پنجره‌گذاری، تعداد گره‌های Random Treeو ... پیشنهاد می‌شود با استفاده از روش‌های بهینه سازی چند پارامتری، به بهینه سازی این پارامترها پرداخته شود.

در پایان می‌توان این نکته را یادآور نمود که دسته‌بندی هر قسمت از دیواره‌ی چاه به سه کلاس به معنای آن است که ما اولاً سعی در کشف شکستگی‌ها داریم و دوماً سعی در پی بردن به این موضوع داریم که آیا شکستگی دارای ویژگی پرشدگی هست یا خیر. این مسأله می‌تواند یک بخش از یک پروژه‌ی بسیار بزرگ یعنی پی بردن به تمام ویژگی‌های شکستگی‌ها باشد.

بنابراین بعنوان فازهای بعدی این پژوهش می‌توان به بررسی سایر ویژگی‌های شکستگی‌ها از قبیل شیب، جهت و میزان بازشدگی پرداخت که بعنوان کارهای آینده می‌توانند بسیار ارزشمند باشند.

منابع و مراجع

[۱] م. رضائی, *زمین شناسی نفت*. تهران: علوی, ۱۳۸۰.

[۲] L. B. Thompson, “Fractured Reservoirs: Integration Is the Key to Optimization,” *J. Pet. Technol.*, vol. ۵۲, no. ۰۲, pp. ۵۲–۵٤, Apr. ۲۰۱۳.

[۳] R. A. Nelson, *Geologic Analysis of Naturally Fractured Reservoirs (Second edition)*. Gulf Professional Publishing, U.S.A., ۲۰۰۱, p. ۳۳۲.

[٤] L. P. Martinez-Torres, “Characterization of naturally fractured reservoirs from conventional well logs,” University of Oklahoma, ۲۰۰۲.

[٥] P. Dutta, C. Mardi, M. Akbar, S. K. Singh, J. Al-Genai, and A. Akhtar, “A Novel Approach To Fracture Characterization Utilizing Borehole Seismic Data,” in *SPE Middle East Oil and Gas Show and Conference*, ۲۰۱۳.

[٦] B. Tokhmechi, H. Memarian, H. a Noubari, and B. Moshiri, “A novel approach proposed for fractured zone detection using petrophysical logs,” *J. Geophys. Eng.*, vol. ٦, no. ٤, pp. ۳٦٥–۳۷۳, Dec. ۲۰۰۹.

[۷] B. Tokhmechi, H. Memarian, and M. R. Rezaee, “Estimation of the fracture density in fractured zones using petrophysical logs,” *J. Pet. Sci. Eng.*, vol. ۷۲, no. ۱–۲, pp. ۲۰٦–۲۱۳, May ۲۰۱۰.

[۸] O. Serra, *Formation Microscanner Image Interpretation*. Houston: Schlumberger Education Services, ۱۹۸۹, p. ۱۱۷.

[۹] R. Aguilera, “Geologic and Engineering Aspects of Naturally Fractured Reservoirs,” *CSEG Rec.*, vol. ۲۸, pp. ٤٥–٤۹, ۲۰۰۳.

[۱۰] A. M. Zellou and A. Ouenes, “Soft Computing and Intelligent Data Analysis in Oil Exploration,” *Dev. Pet. Sci.*, vol. ٥۱, pp. ٥۸۳–٦۰۲, ۲۰۰۳.

[۱۱] M. Nemati and H. Pezeshk, “Spatial Distribution of Fractures in the Asmari Formation of Iran in Subsurface Environment: Effect of Lithology and Petrophysical Properties,” *Nat. Resour. Res.*, vol. ۱٤, no. ٤, pp. ۳۰٥–۳۱٦, Mar. ۲۰۰٦.

[۱۲] O. P. Wennberg, T. Svana, M. Azizzadeh, A. M. M. Aqrawi, P. Brockbank, K. B. Lyslo, and S. Ogilvie, “Fracture intensity vs. mechanical stratigraphy in platform top carbonates: the Aquitanian of the Asmari Formation, Khaviz Anticline, Zagros, SW Iran,” *Pet. Geosci.*, vol. ۱۲, no. ۳, pp. ۲۳٥–۲٤٦, Aug. ۲۰۰٦.

[۱۳] D. a. Sagi, M. Arnhild, and J. F. Karlo, “Quantifying fracture density and connectivity of fractured chalk reservoirs from core samples: implications for fluid flow,” *Geol. Soc. London, Spec. Publ.*, vol. ۳۷٤, ۲۰۱۳.

[١٤] A. Mohebbi, M. Haghighi, and M. Sahimi, “Conventional Logs for Fracture Detection & Characterization in One of the Iranian Field,” in *International Petroleum Technology Conference*, ٢٠٠٧.

[١٥] B. J. Stephenson, A. Koopman, H. Hillgartner, H. McQuillan, S. Bourne, J. J. Noad, and K. Rawnsley, “Structural and stratigraphic controls on fold-related fracturing in the Zagros Mountains, Iran: implications for reservoir development,” *Geol. Soc. London, Spec. Publ.*, vol. ٢٧٠, no. ١, pp. ١–٢١, Sep. ٢٠١٣.

[١٦] B. Tokhmechi, H. Memarian, V. Rasouli, H. A. Noubari, and B. Moshiri, “Fracture detection from water saturation log data using a Fourier–wavelet approach,” *J. Pet. Sci. Eng.*, vol. ٦٩, no. ١–٢, pp. ١٢٩–١٣٨, Nov. ٢٠٠٩.

[١٧] H. Soroush, V. Rasouli, and B. Tokhmechi, “A combined Bayesian–wavelet–data fusion approach for borehole enlargement identification in carbonates,” *Int. J. Rock Mech. Min. Sci.*, vol. ٤٧, no. ٦, pp. ٩٩٦–١٠٠٥, Sep. ٢٠١٠.

[١٨] H. Soroush, V. Rasouli, and B. Tokhmechi, “A data processing algorithm proposed for identification of breakout zones in tight formations: A case study in Barnett gas shale,” *J. Pet. Sci. Eng.*, vol. ٧٤, no. ٣–٤, pp. ١٥٤–١٦٢, Nov. ٢٠١٠.

[١٩] X.-F. Zhang, B.-Z. Pan, F. Wang, and X. Han, “A study of wavelet transforms applied for fracture identification and fracture density evaluation,” *Appl. Geophys.*, vol. ٨, no. ٢, pp. ١٦٤–١٦٩, Jul. ٢٠١١.

[٢٠] P. Masoudi, B. Tokhmechi, A. Bashari, and M. A. Jafari, “Identifying productive zones of the Sarvak formation by integrating outputs of different classification methods,” *J. Geophys. Eng.*, vol. ٩, no. ٣, pp. ٢٨٢–٢٩٠, Jun. ٢٠١٢.

[٢١] P. Masoudi, B. Tokhmechi, M. A. Jafari, and B. Moshiri, “Application of fuzzy classifier fusion in determining productive zones in oil wells,” *Energy, Explor. Exploit.*, vol. ٣٠, no. ٣, pp. ٤٠٣–٤١٦, Jun. ٢٠١٢.

[٢٢] A. A. Zerrouki, T. Aïfa, and K. Baddari, “Prediction of natural fracture porosity from well log data by means of fuzzy ranking and an artificial neural network in Hassi Messaoud oil field, Algeria,” *J. Pet. Sci. Eng.*, vol. ١١٥, pp. ٧٨–٨٩, Mar. ٢٠١٤.

[٢٣] P. Masoudi, Y. Asgarinezhad, and B. Tokhmechi, “Feature selection for reservoir characterisation by Bayesian network,” *Arab. J. Geosci.*, Apr. ٢٠١٤.

[٢٤] A. Mollajan, Y. Asgarinezhad, B. Tokhmechi, and S. Sherkati, “A comparative study of two data-driven methods in detection of vuggy zones: a case study from a carbonate reservoir, west of Iran,” *Carbonates and Evaporites*, Jun. ٢٠١٤.

[٢٥] Y. Asgarinezhad, B. Tokhmechi, A. Kamkar Rouhani, S. Sherkati, K. Kavousi, and A. Jamali, “A combined Parzen-wavelet approach for detection of vuggy

zones in fractured carbonate reservoirs using petrophysical logs," *J. Pet. Sci. Eng.*, vol. ۱۱۹, pp. ۱–۷, Jul. ۲۰۱٤.

[۲٦] ح. معماریان و ب. تخمه چی, "شناسایی شکستگی ها از چاه نمودارهای پتروفیزیکی. کارفرما شرکت ملی نفت مناطق نفت خیز جنوب," ۱۳۹۲.

[۲۷] S. Luthi, *Geological Well Logs: Their Use in Reservoir Modeling*. Springer Science & Business Media, ۲۰۰۱, p. ۳۷۳.

[۲۸] F. Khoshbakht, "Application of borehole image logs in fracture study in one of oil field of south west Iran," Tehran University, Tehran, ۲۰۰٦.

[۲۹] S. E. Prensky, "Advances in borehole imaging technology and applications," *Geol. Soc. London, Spec. Publ.*, vol. ۱٥۹, no. ۱, pp. ۱–٤۳, Jan. ۱۹۹۹.

[۳۰] M. Seifallahi, "Intelligent Detection of Fractures in Carbonate Reservoirs Using Image Logs," Shahrood University, Shahroud, ۲۰۱۲.

[۳۱] A. Bifet and R. Gavalda, "Learning from Time-Changing Data with Adaptive Windowing," in *Society for Industrial and Applied Mathematics*, ۲۰۰۷, p. ٦.

[۳۲] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, ۲۰۱۳.

[۳۳] A. Phinyomark, P. Phukpattaranont, and C. Limsakul, "Feature reduction and selection for EMG signal classification," *Expert Syst. Appl.*, vol. ۳۹, no. ۸, pp. ۷٤۲۰–۷٤۳۱, Jun. ۲۰۱۲.

[۳٤] M. Akay, *Biomedical signal processing*. Academic Press, ۲۰۱۲.

[۳٥] G. Casella and R. L. Berger, *Statistical inference*, vol. ۲. Duxbury Pacific Grove, CA, ۲۰۰۲.

[۳٦] K. P. Balanda and H. L. Macgillivray, "Kurtosis: A Critical Review," *Am. Stat.*, vol. ٤۲, no. ۲, pp. ۱۱۱–۱۱۹, May ۱۹۸۸.

[۳۷] B. M. Bolstad, R. . Irizarry, M. Astrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, vol. ۱۹, no. ۲, pp. ۱۸٥–۱۹۳, Jan. ۲۰۰۳.

[۳۸] R. Setiono, "Chi۲: feature selection and discretization of numeric attributes," in *Proceedings of ۷th IEEE International Conference on Tools with Artificial Intelligence*, pp. ۳۸۸–۳۹۱.

[۳۹] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mob. Comput. Commun. Rev.*, vol. ٥, no. ۱, p. ۳, Jan. ۲۰۰۱.

[٤۰] D. J. C. Mackay, *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press, ۲۰۰۳.

[٤١] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, ٢٠٠٨.

[٤٢] M. I. J. Andrew Y. Ng, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes," in *Neural Information Processing Systems*, ٢٠٠٢, pp. ٨٤١–٨٤٨.

[٤٣] G. Forman and I. Cohen, "Learning from Little: Comparison of Classifiers Given Little Training," *Knowl. Discov. Databases PKDD*, vol. ٣٢٠٢, pp. ١٦١–١٧٢, ٢٠٠٤.

[٤٤] M. Ramoni and P. Sebastiani, "Robust Bayes classifiers," *Artif. Intell.*, vol. ١٢٥, no. ١–٢, pp. ٢٠٩–٢٢٦, Jan. ٢٠٠١.

[٤٥] J. A. Hartigan, *Bayes Theory*. New York, NY: Springer New York, ١٩٨٣.

[٤٦] T. M. Mitchell, *Machine Learning*. McGrawHill Science, ١٩٩٧.

[٤٧] R. Rastogi and K. Shim, "PUBLIC: a decision tree classifier that integrates building and pruning," in *VLDB*, ١٩٩٨, vol. ٩٨, pp. ٢٤–٢٧.

[٤٨] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. CRC Press, ٢٠٠٩, p. ٢٠٨.

[٤٩] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, ٢٠٠٤, p. ٣٠٠.

[٥٠] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Stat. Surv.*, vol. ٤, pp. ٤٠–٧٩, ٢٠١٠.

Abstract

Almost ۵۰ percent of oil reservoirs through the world are naturally fractured and in Iran this ratio is up to ۹۰ percent. While accurate fractures determining of an oil field is a very huge work which needs step by step documentation and description of huge number of image well logs by experts. In addition already and especially in Iran there are many obstacles in the way of preparing image well logs, this problem have made it unavoidable using other datasets for detecting fractures.

Until now many different studies have been developed in order of separating between open and close natural fractures but none of them achieved the end goal of usability in the industry. In this study it's tried to use petrophysical data for finding out if a zone of the well is not fractured, open fractured or close fractured.

Suggested method uses pre-processed petrophysical data instead of raw data for data mining. This pre-processing includes windowing method following with replacing the data with the variance of windowed sub-data-signal. After pre-processing and for data classification, we use two classifiers which finally a majority-voting algorithm ensembles their answers.

We experienced many different classification algorithms and at last Naïve Bayes and Random Tree which showed the most correct classification results have been selected to be used in this regard.

This study has been developed on some oil fields of Gachsaran oil field in south west of Iran. Because of some shortages in the data, we proposed two series of experiments, one of them including the data from ۷ wells with ۷ well logs and the other one including ۴ wells with ۱۱ well logs.

Although the main approach of this study aims to data classification into three classes of open, close and not fractured zones, but also in the last section we proposed our suggested method for classifying the data into two classes of fractured and not-fractured zones, so other researchers can compare their two class classifying results with the suggested method.

Throughout the study one of the main principles is to establishing balance between accuracy of three classes. But unfortunately this is not compatible with the nature of natural fractures, because these three classes have no balance and not-fractured is the dominant class. So we are constrained to accept only well balanced classification results which indeed are not the most accurate results.

In conclusion, final results shows the accuracy of ۷۷,۵ percent for two classes and more than ۸۰,۶ for three classes classification in the best cases. Also the mean of accuracy for two classes is more than ۷۰,۱ and for three classes is more than ۶۸,۶ percent.

Key Words:

Data mining, fracture, open fracture, close fracture, classification, Naïve Bayes, Random Tree, Majority Voting, Petrophysical well-log, Image well-log, Windowing.

Shahrood University

Faculty of Computer Engineering & IT

Intelligent detection of fracture specifications in petro-physic well logs

Amir Keshavarzreza

Supervisors:

Dr. Hamid Hassanpour

Dr. Behzad Tokhmchi

Associate Supervisor:

Mohamad Hossein Khosravi

۲۰۱۴