ساخت نگاره گرافیکی مصنوعی چاه توسط روش یادگیری حساس‌به‌هزینه

محمد صالح جمشیدی گوهری[۱]، محمد امامی نیری[۲]، جواد قیاسی فریز[۳]، سعید صادق نژاد[۴]
[۱] دانشجوی دکتری، دانشگاه تهران، saleh.jamshidi@ut.ac.ir
[۲] استادیار، دانشکده مهندسی شیمی، دانشکده فنی، دانشگاه تهران، emami.m@ut.ac.ir
[۳] دکتری مهندسی نفت، شرکت نفت مناطق مرکزی ایران، javadghiasi86@gmail.com
[۴] استادیار، دانشگاه تربیت مدرس، sadeghnejad@modares.ac.ir

چکیده

نگاره گرافیکی چاه، تصویری بصری سریع از وضعیت زمین‌شناسی و رخساره‌های زیرِ زمین به دست می‌دهد. در صورت عدم دسترسی به نگاره گرافیکی یک چاه، می‌توان با استفاده از روش‌های یادگیری ماشین، رخساره‌ها را از روی‌داده‌های نگاره‌ها و یا مغزه چاه‌ها پیش‌بینی کرد. در این راستا، چالشِ عدم تعادل مابین کلاس‌ها، یادگیری کافی و تمایز بین کلاس‌های اقلیت و اکثریت را برای طبقه‌بندی کننده‌ها دشوار می‌سازد. این تحقیق بر توسعه یک روش طبقه‌بندی کننده حساس‌به‌هزینه با اصلاح تابع هدف الگوریتم متداول جنگل تصادفی به‌منظور ساخت نگاره گرافیکی مصنوعی چاه تمرکز دارد. ابتدا، نسخه استاندارد این الگوریتم پیاده‌سازی شد تا پایه‌ای برای مقایسه عملکرد ارائه شود. در مرحله دوم، الگوریتم حساس‌به‌هزینه پیشنهادی توسعه یافت. نتایج نشان می‌دهد که روش حساس به هزینه عملکرد بهتری نسبت به الگوریتم استاندارد در مسئله موردنظر این پژوهش دارد.

واژه‌های کلیدی: یادگیری حساس‌به‌هزینه، نگاره گرافیکی چاه، طبقه‌بندی چند کلاسه، عدم تعادل کلاس، جنگل تصادفی، نگاره چاه

Synthetic graphic well logs generation via Cost-Sensitive Learning

Mohammad Saleh Jamshidi Gohari [1], Mohammad Emami Niri [2], Javad Ghiasi-Freez [3], Saeid Sadeghnejad[4]
[1] PhD Student, University of Tehran
[2] Assistant Professor, College of Engineering, University of Tehran
[3] PhD in Petroleum Engineering, Central Iranian Oil Company
[4] Assistant Professor, Tarbiat Modares University

Abstract

The graphic well log offers a quick sight view of the geological condition and underground facies. If it isn't viable to access the graphic well log, machine learning techniques may be used to predict facies from well logs or cores. In this regard, the class imbalance venture makes it difficult for classifiers to adequately learn and distinguish among minority and majority classes. This research focuses on the development of a cost-sensitive classifier by modifying the objective function of the well-known random forest algorithm to construct a synthetic graphic well log. First, a standard version of the algorithm was implemented to provide a basis for performance comparison. In the second stage, a cost-sensitive algorithm was developed. The results indicate that the cost-sensitive method performs better than the standard algorithm in this case study.

Keywords: Cost-Sensitive Learning, Graphic Well Log, Multiclass Classification, Class Imbalance, Random Forest, Well Log

۱ مقدمه

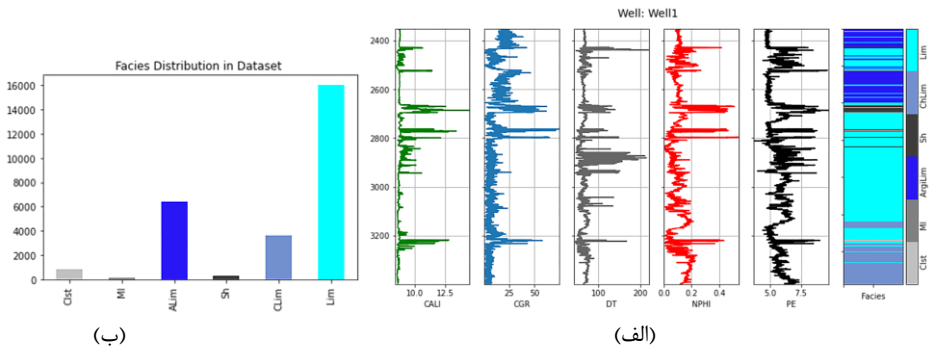
روش استاندارد برای جمع‌آوری سریع اطلاعات از چاه در حال حفاری، ساخت یک نگاره گرافیکی از توالی و سکانس در مسیر حفاری داخل چاه است. نگاره‌های گرافیکی چاه، تصویری بصری سریع از مقطع زیرِ زمین ایجاد می‌کنند و روشی مناسب برای ایجاد همبستگی و مقایسه بین مقاطع یا بخش‌های معادل از مناطق مختلف هستند. یکی از چالش‌های موجود در برخی از میدان‌ها عدم دسترسی به نگاره گرافیکی چاه یا در مواردی اطلاعات تفسیر شده سنگ‌شناسی در برخی از چاه‌های میدان است. در حال حاضر به کمک روش‌های مختلف یادگیریِ ماشین می‌توان رخساره‌ها را از روی‌داده‌های نگاره و مغزه چاه‌های مجاور پیش‌بینی کرد. دوبوا، و همکاران (۲۰۰۷) چهار روش مختلف قانون بیز، منطق فازی، نزدیک‌ترین همسایه کِی و شبکه عصبی پس انتشار پیش‌خور را در یک مسئله طبقه‌بندی رخساره سنگ موردبررسی قراردادند. بر اساس نتایج حاصله در میدان موردمطالعه شبکه عصبی مصنوعی نسبت به سایر طبقه‌بندی کننده‌ها عملکرد بهتری داشت. لیندبرگ و گرانا (۲۰۱۵) جهت طبقه‌بندی رخساره‌ها یک روش طبقه‌بندی مبتنی بر مدل‌های مارکوف پنهان پیشنهاد کردند. نتایج حاصله، توانایی روش پیشنهادی را در استنباط پارامترهای مدل مانند احتمالات انتقال و توزیع‌های پسین رخساره را نشان می‌دهد. بستاجینی و همکاران (۲۰۱۷) از یک پایپلاین یادگیری ماشین برای طبقه‌بندی رخساره‌ها بر اساس اندازه‌گیری‌های نگاره استفاده کردند. بر اساس نتایج آن‌ها افزایش ویژگی در الگوریتم پیشنهادی منجر به افزایش دقت می‌شود. زی و اسپایکس (۲۰۲۲) تکنیکی برای طبقه‌بندی رخساره‌های سنگی از داده‌های نگاره چاه با استفاده از یک الگوریتم یادگیری نیمه نظارتی ارائه کردند. نتایج طبقه‌بندی نشان داد که رخساره‌های پیش‌بینی‌شده با روش یادگیری نیمه نظارتی، مطابقت خوبی با برچسب‌های واقعی دارند.

همان‌گونه در بالا اشاره شد تاکنون از الگوریتم‌های یادگیری ماشین مختلف جهت تخمین رخساره‌های سنگی از روی‌داده‌های نگاره چاه و مغزه استفاده‌شده است. در اینجا هدف ارائه یک مدل یادگیری ماشین بهینه جهت ساخت نگاره گرافیکی مصنوعی چاه در چاه‌هایی است که در آن‌ها دسترسی به آن به‌صورت کامل یا مقطعی وجود ندارد.

منطقه موردمطالعه، یک از میادین نفتی جنوب غربی ایران است؛ که ازنظر جغرافیایی در استان خوزستان و در نزدیکی مرز ایران و عراق و ازنظر زمین‌شناسی، در کمربند انتقالی بین منطقه چین‌خورده زاگرس و حاشیه پایدار شرقی پلتفرم عربی قرار دارد.

۲ روش تحقیق

داده‌های معرفی‌شده در این مطالعه از داده‌های پتروفیزیکی مرسوم، شامل نگاره‌های پرتوگاما، تخلخل نوترون، فاکتور فوتوالکتریک، چگالی، سونیک، قطر سنجی و همچنین داده‌های نگاره گرافیکیِ چاه چهار چاه (با ۳۴۱۴۸ نمونه) از میدان نفتی موردمطالعه جمع‌آوری‌شده است. با استفاده از نگاره گرافیکی چاه داده‌های نگاره به‌صورت دستی برای هر عمق، برچسب‌گذاری شدند (شکل ۱(الف)).



شکل ۱. (الف) نمودار نگاره‌ها برای چاه شماره ۱(به‌عنوان نمونه) و (ب) توزیع رخساره (کلاس) در مجموعه داده.

با توجه به موقعیت جغرافیایی و شمول طیف داده‌های موجود در چاه‌ها یک چاه به‌عنوان چاه کور انتخاب شد. در فرآیند

پیش‌پردازش داده‌ها، متغیرهای ستونی که فقط یک مقدار واحد داشتند، متغیرهای ستونی با مقادیر بسیار کم، ستون‌های دارای واریانس کم و سطرهای حاوی مشاهدات تکراری و مقدار منفی، شناسایی و حذف شدند. سپس با استفاده از ایجاد همبستگی بین ویژگی‌های مختلف، ستون‌های غیرضروری نیز حذف گردیدند. با توجه به توزیع کمیت، از استانداردسازی استفاده شد و داده‌ها استاندارد شدند. یکی از اهداف در مرحله پیش‌پردازش دستیابی به حد بالایی از میزان جداسازی بین کلاس‌ها در فاز قبل از آموزش یا آزمون طبقه‌بندی کننده است. برای این منظور از آنالیز تشخیص خطی با دقت ۸۵ درصد در مرحله پیش‌پردازش در فاز قبل از آموزش استفاده شد. تکنیک‌های مختلفی ازجمله جمع‌آوری داده‌های بیشتر، تغییر معیار عملکرد، نمونه‌گیری مجدد از مجموعه داده، تولید نمونه‌های مصنوعی، امتحان الگوریتم‌های مختلف، مدل‌های تاوانیدِ و غیره برای مواجهه با مشکل داده‌های آموزشی نامتعادل (شکل ۱(ب)) وجود دارد. در اکثر الگوریتم‌های یادگیری ماشین فرض می‌شود که تعداد نمونه‌ها در همه کلاس‌ها یکسان است.

این الگوریتم‌ها را باید به‌گونه‌ای اصلاح کرد که یادگیری به سمت کلاس‌هایی که نمونه کمتری در داده‌های آموزشی دارند سوگیری شود. در این تحقیق تمرکز بر توسعه طبقه‌بندی کننده حساس به هزینه با اصلاح تابع هدف الگوریتم معروف، جنگل تصادفی است. درحالی‌که جنگل تصادفی برای کاربردهای متعدد مناسب است، عملکرد ضعیفی در وظایف طبقه‌بندی نامتعادل دارد. برای تغییر جنگل تصادفی استاندارد، جهت حساسیت به هزینه، وزن‌هایی به کلاس‌های مختلف اختصاص داده شد. از معکوس توزیع کلاس استفاده شد و درنتیجه الگوریتم مجبور شد تا روی کلاس اقلیت تمرکز بیشتری داشته باشد. نتایج حاصله با یک الگوریتم جنگل تصادفی استاندارد مقایسه شد. برای ارزیابی عملکرد الگوریتم‌ها در این مقاله، از برخی معیارهای ارزیابی مانند دقت، صحت، بازیابی، معیار اف، ضریب کاپا کوهن و اندیس ژاکار استفاده شد. در یادگیری ماشین می‌توان از ضریب کاپا کوهن به‌عنوان یک معیار از عملکرد طبقه‌بندی کننده استفاده کرد.

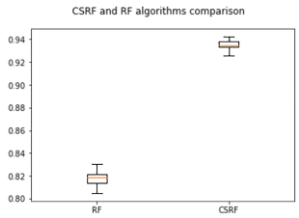
$$Kappa = \frac{p_0 - p_c}{1 - p_c} \qquad (۱)$$

که در این رابطه p_0 نشان‌دهنده توافق نسبی مشاهده‌شده در بین طبقه‌بندی کننده‌ها و p_c نشان‌دهنده احتمال این است که این توافق به دلیل شانس باشد (مینی و سان (۲۰۲۱)). اندیس ژاکار یا ضریب شباهت ژاکار معیاری برای مقایسه شباهت یا تفاوت مجموعه نمونه‌های آماری است.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad (۲)$$

که در این رابطه A مجموعه نمونه اول، B مجموعه نمونه دوم و $J(A,B)$ اندیس ژاکار است (فلچر و اسلام (۲۰۱۸)).

جهت سفارشی کردن مدل برای داده‌ها، هایپر پارامترها تنظیم گردیدند. شکل ۲(الف) یک نمودار جعبه و خط را ارائه می‌دهد که گسترش امتیازهای دقت را در هر فولد (Fold) اعتبارسنجی متقابل برای الگوریتم‌های جنگل تصادفی حساس به هزینه (CSRFC) و جنگل تصادفی استاندارد (RFC) نشان می‌دهد. ماتریس درهم‌ریختگی طبقه‌بندی کننده‌ها برای خروجی آزمون در شکل ۲ (ب و ج) نشان داده‌شده است.



(ج)

Pred / True	Clst	Ml	ALim	Sh	CLim	Lim	Total
Clst	161	2	4	1	1	6	175
Ml	12	12	1		1		26
ALim	3		1172			125	1300
Sh	14		2	45		13	74
CLim			1		677	98	776
Lim	1		49		29	3065	3144
Precision	0.84	0.86	0.95	0.98	0.96	0.93	0.94
Recall	0.92	0.46	0.90	0.61	0.87	0.97	0.93
F1	0.88	0.60	0.93	0.75	0.91	0.95	0.93

(ب)

Pred / True	Clst	Ml	ALim	Sh	CLim	Lim	Total
Clst	125		5		15	30	175
Ml	21					5	26
ALim	1		999			300	1300
Sh	41		1			32	74
CLim			1		407	368	776
Lim	1		83		88	2972	3144
Precision	0.66	0.00	0.92	0.00	0.80	0.80	0.81
Recall	0.71	0.00	0.77	0.00	0.52	0.95	0.82
F1	0.69	0.00	0.84	0.00	0.63	0.87	0.81

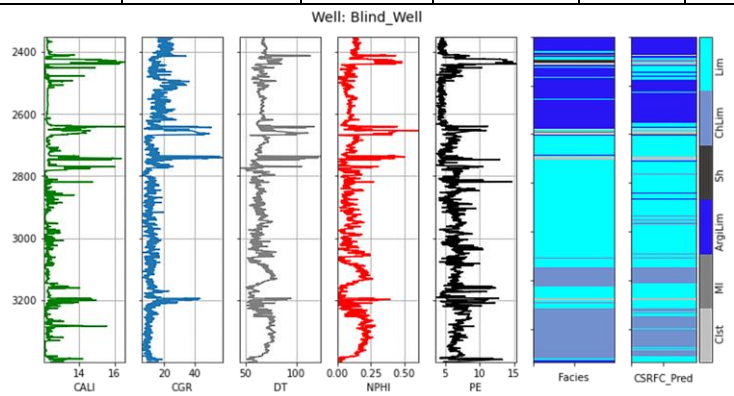
(الف)

شکل ۲. (الف) نمودار جعبه و خط برای الگوریتم‌های (CSRFC) و (RFC). (ب) ماتریس درهم‌ریختگی (CSRFC) برای داده‌های آزمون. (ج) ماتریس درهم‌ریختگی (RFC) برای داده‌های آزمون.

نتایج ارزیابی عملکرد طبقه‌بندی کننده‌ها توسط معیارهای مختلف بر روی‌داده‌های آزمون و چاه کور در جدول ۱ آورده شده است. در شکل ۳ نگاره گرافیکی چاه با استفاده از طبقه‌بندی کننده بهینه (CSRFC) برای داده‌های چاه کور رسم شده است.

جدول ۱. نتایج ارزیابی عملکرد طبقه‌بندی کننده‌ها بر روی‌داده‌های آزمون و چاه کور (نتایج مدل بهینه با حروف پررنگ مشخص‌شده است).

Data type	Model type	Jaccard	F1-Score	Accuracy	Recall	Precision	Kappa
Test	**CSRFC**	**۰/۹۰**	**۰/۹۳**	**۰/۹۳**	**۰/۹۳**	**۰/۹۴**	**۰/۹۰**
	RFC	۰/۶۹	۰/۸۰	۰/۸۱	۰/۸۱	۰/۸۰	۰/۶۷
Blind	**CSRFC**	**۰/۷۱**	**۰/۸۲**	**۰/۸۳**	**۰/۸۳**	**۰/۸۲**	**۰/۷۳**
	RFC	۰/۵۵	۰/۷۰	۰/۷۲	۰/۷۲	۰/۷۱	۰/۵۲



شکل ۳. نگاره گرافیکی چاه تولیدشده با استفاده از (CSRFC) برای داده‌های چاه کور (Facies نگاره گرافیکی چاه واقعی است).

۳ نتیجه‌گیری

در این مقاله از یادگیری حساس به هزینه جهت آموزش الگوریتم جنگل تصادفی باهدف ساخت نگاره گرافیکی مصنوعی استفاده شد. یادگیری حساس به هزینه زیرشاخه‌ای از یادگیری ماشین است که هزینه‌های خطاهای پیش‌بینی را هنگام آموزش مدل در نظر می‌گیرد. این‌یک زمینه مطالعاتی است که ارتباط نزدیکی با حوزه یادگیری نامتعادل چند کلاسه موجود در این تحقیق دارد. علاوه بر این، برای وزن دهی کلاس از ابتکار معکوس توزیع کلاس در مجموعه داده استفاده شد. نتایج حاصله حاکی از آن است که بر اساس معیارهای ارزیابی مختلف (جدول ۱) طبقه‌بندی کننده حساس به هزینه جنگل تصادفی همراه با روش وزن دهی معکوس توزیع کلاس، جهت ساخت نگاره گرافیکی چاه مصنوعی در مقایسه با الگوریتم استاندارد عملکرد بهتری دارد. با توجه به ضریب شباهت ژاکار نگاره گرافیکی چاه مصنوعی تولیدشده با استفاده از (CSRFC) (شکل ۳) ۷۱ درصد (در مقابل ۵۵ درصد برای (RFC)) با نمونه واقعی مطابقت دارد.

منابع

Bestagini, P., V. Lipari and S. Tubaro (2017). A machine learning approach to facies classification using well logs. Seg technical program expanded abstracts 2017, Society of Exploration Geophysicists**:** 2137-2142.

Dubois, M. K., G. C. Bohling and S. Chakrabarti (2007). "Comparison of four approaches to a rock facies classification problem." Computers & Geosciences **33**(5): 599-617.

Fletcher, S. and M. Z. Islam (2018). "Comparing sets of patterns with the Jaccard index." Australasian Journal of Information Systems **22**.

Lindberg, D. V. and D. Grana (2015). "Petro-elastic log-facies classification using the expectation–maximization algorithm and hidden Markov models." Mathematical Geosciences **47**(6): 719-752.

Mienye, I. D. and Y. Sun (2021). "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data." Informatics in Medicine Unlocked **25**: 100690.

Xie, W. and K. T. Spikes (2022). "Well-log facies classification using an active semi-supervised algorithm with pairwise constraints." Geophysical Journal International **229**(1): 56-69.