مجله ژئوفیزیک ایران، جلد ۱۹، شماره ۲، ۱۴۰۴، صفحه ۸۱–۱۰۰ (مقاله پژوهشی) DOI: 10.30499/IJG.2024.450092.1589

کاهش داده‌های مورد نیاز برای آموزش مدل‌های یادگیری عمیق بر اساس خوشه‌بندی داده‌ها و کاربرد آن در وارون‌سازی یک‌بعدی مگنتوتلوریک

مهدی رحمانی‌جوینانی[1] و بنفشه حبیبیان‌دهکردی[2]*

[1] دانشجوی دکترای، موسسه ژئوفیزیک دانشگاه تهران، تهران، ایران
[2] استادیار، موسسه ژئوفیزیک دانشگاه تهران، تهران، ایران

(دریافت: ۱۴۰۳/۰۱/۰۸، پذیرش: ۱۴۰۳/۰۷/۰۴)

چکیده

رویکردهای یادگیری عمیق داده‌محور با چالش تولید داده‌هایی به تعداد زیاد و با کیفیت بالا و بار محاسباتی سنگین و زمان آموزش طولانی تحمیل شده توسط آن روبرو هستند. علاوه بر این، در صورتی‌که بعد از جداسازی تصادفی داده‌ها به سه مجموعه آموزش، اعتبارسنجی و آزمایش، توزیع آماری یکسانی برای آنها به‌دست نیاید، به دلیل رفتار نامنظم منحنی خطای آموزش و اعتبارسنجی، تعمیم‌پذیری خوبی حاصل نمی‌شود. در این پژوهش با استفاده از رویکرد مبتنی بر خوشه‌بندی اولیه داده‌ها و اختصاص درصد مشخصی از هر خوشه به سه مجموعه، و با پیمایش نتایج پیش‌بینی، کمینه داده مورد نیاز برای وارون‌سازی با رویکرد یادگیری عمیق ارائه می‌گردد. با اعمال آزمون‌های آماری نشان داده می‌شود که داده‌هایی که با این رویکرد جداسازی شده‌اند، دارای توزیع یکسان در سه مجموعه هستند. یک مدل یادگیری عمیق مبتنی بر معماری U-Net برای وارون‌سازی یک‌بعدی داده‌های مگنتوتلوریک آموزش داده می‌شود. به این منظور از یک مدل ژئوالکتریکی پنج لایه که شرایط یک میدان زمین‌گرمایی را شبیه‌سازی می‌کند، استفاده شده‌است. آموزش شبکه با تعداد متفاوت داده‌هایی که با روش گفته شده جداسازی شده‌اند، تکرار و عملکرد آن با معیارهای کمی و کیفی متفاوتی سنجیده می‌شود. با پیمایش نتایج وارون سازی با داده‌های آزمایشی یکسان بر مدل‌های آموزش دیده با درصد داده‌ای مختلف می‌توان بدون اینکه از دقت شبکه کاسته شود، به میزان ۵۰ درصد تعداد داده‌های مورد نیاز برای آموزش مدل یادگیری عمیق و بنابراین زمان آموزش را کاهش داد. در مواجهه با داده‌های پیچیده‌تر، واقعی‌تر و نویزی قطعا جداسازی تصادفی رهیافت مناسبی برای تشکیل سه مجموعه نیست. هرچه شرایط پیچیده‌تر و تعداد ویژگی‌ها بیشتر باشد، جداسازی تصادفی راهکار نامناسب‌تری است؛ چراکه تفاوت توزیع‌های آماری سه مجموعه بیشتر می‌شود؛ و در نتیجه تعمیم‌پذیری کاهش و تعداد داده‌های مورد نیاز افزایش می‌یابد. در این‌صورت استفاده از خوشه‌بندی راهکار مناسبی برای یکسان‌سازی توزیع آماری سه مجموعه و کاهش تعداد داده‌هاست.

کلمه‌های کلیدی: خوشه‌بندی، مگنتوتلوریک، یادگیری عمیق، وارون‌سازی

* نگارنده رابط: bhabibian@ut.ac.ir

۱ مقدمه

در سال‌های اخیر، نقش روش‌های مبتنی بر هوش مصنوعی و به‌ویژه یادگیری عمیق در جنبه‌های مختلف کار با داده‌های ژئوفیزیکی بسیار پررنگ شده‌است. روش مگنتوتلوریک، یک روش ژئوفیزیکی الکترومغناطیسی با کاربردهای متنوع در زمینه‌هایی همچون مطالعات تکتونیکی (برای مثال کامیو و همکاران، ۲۰۲۲) و اکتشاف منابع زمین‌گرمایی و هیدروکربنی (برای مثال سگوویا و همکاران، ۲۰۲۱؛ میری و همکاران، ۲۰۲۱) است. برای وارون‌سازی و تفسیر سنتی سونداژهای MT، متداول‌ترین روش‌های مورد استفاده توسط کانستبل و همکاران (۱۹۸۷)، پارکر و بوکر (۱۹۹۶)، اسمیت و بوکر (۱۹۸۸) و فیشر و همکاران (۱۹۸۱) ارائه شده‌اند. در پژوهش‌های بعدی از تانسور فاز (کالدول وهمکاران، ۲۰۰۴) برای وارون‌سازی یک‌بعدی مگنتوتلوریک استفاده شده‌است (یونگه، ۲۰۱۱). العلی و همکاران (۲۰۲۰) رهیافت جدیدی برای تعیین تعداد بهینه لایه‌ها ارائه دادند. پارامترهای مدل یک‌بعدی شامل مقاومت ویژه الکتریکی و ضخامت لایه‌ها در قالب یک مسئله وارون بدوضع از روی داده‌ها تعیین می‌شوند. روش‌های یادگیری عمیق به‌عنوان رهیافت‌های آماری مبتنی بر داده قابل کاربرد برای حل مسائل وارون بدوضع هستند. این وارون‌سازی یک مسئله رگرسیون شدیدا غیرخطی است و شبکه‌های عمیق‌تر توابع غیرخطی را بهتر تقریب می‌زنند. مزیت بهره گرفتن از لایه‌های متعدد در شبکه‌های عمیق، تولید توابع نگاشت غیرخطی با پیچیدگی بالاست. نمونه‌های متعددی از کاربرد روش‌های یادگیری عمیق در مدل‌سازی‌های الکترومغناطیسی (اوه و همکاران، ۲۰۲۰؛ پازیرف، ۲۰۱۹؛ شهریاری و همکاران، ۲۰۲۰) و مگنتوتلوریک (لیو و همکاران، ۲۰۲۱ و ۲۰۲۲؛ لیائو و همکاران، ۲۰۲۱ الف و ب)، گرایش فزاینده به استفاده از این روش‌ها در مدل‌سازی‌های ژئوفیزیکی را نشان می‌دهد.

روش‌های داده‌محور یادگیری عمیق برای یادگیری ویژگی‌های یک سیستم، به داده‌های آموزشی متکی هستند؛ به‌طوری‌که فرایند یادگیری بر اساس تجربه و با شناسایی الگوی پیچیده موجود در داده‌ها محقق می‌شود. هرچه مدل پیچیده‌تر باشد، داده‌های آموزشی بیشتری مورد نیاز است. اگر کمیت و کیفیت نمونه‌ها مناسب نباشد و یا مسئله مورد نظر را به خوبی نمایندگی نکنند، تعمیم‌پذیری و کارآیی مدل تحت شرایط متغیر با مشکل مواجه می‌شود. بار محاسباتی لازم برای تولید این حجم از داده شبیه‌سازی شده برای آموزش بسیار بالاست؛ گرچه مدل آموزش‌دیده در تولید خروجی به سرعت عمل می‌کند. بنابراین چالش اول، تولید داده‌های کافی و با کیفیت بالاست که معرف سناریوهای واقعی باشند. چالش بعدی، آموزش شبکه با این حجم از داده است که مستلزم انجام محاسبات سنگین و صرف زمان طولانی است. چالش پایانی هم تعمیم‌پذیری مدل آموزش داده‌شده است.

اولین گام پس از تولید داده در یادگیری عمیق، جداسازی داده‌ها در قالب سه مجموعه آموزش، اعتبارسنجی و آزمایش است. اگر این جداسازی به‌شکل تصادفی انجام شود، تضمینی وجود ندارد که سه مجموعه داده از توزیع آماری یکسانی برخوردار باشند. در این پژوهش رویکرد خوشه‌بندی داده‌ها و اختصاص درصد مشخصی از هر خوشه به سه مجموعه برای جداسازی به‌کار گرفته می‌شود. بطور خاص از معماری U-Net برای وارون‌سازی یک‌بعدی داده‌های مگنتوتلوریک استفاده شده است. شبکه U-Net بر اساس مطالعه‌ای که در آن سه معماری U-Net، Residual Network (Res-Net) و Variational Autoencoder (VAE) به‌عنوان سه مدل یادگیری عمیق از لحاظ دقت و تعمیم‌پذیری در وارون‌سازی داده‌های MT مقایسه شده‌اند (رحمانی جوینانی و همکاران، ۲۰۲۴)، انتخاب شده‌است. به این منظور در مطالعه مورد اشاره، معیارهای متنوعی به‌کار گرفته

شدند و عملکرد معماری‌های متفاوت به‌طور عددی و تحلیلی مورد ارزیابی قرار گرفتند. نتایج حاصل، نشان‌دهنده برتری نسبی معماری U/Net در بازیابی مدل‌های لایه‌ای مقاومت ویژه الکتریکی بود. این مطالعه مقایسه‌ای را می‌توان به‌طور خاص مبنایی برای مدل‌سازی MT با رهیافت یادگیری عمیق و بررسی قابلیت‌های آن تلقی کرد. ضمنا ضخامت لایه‌ها متغیر و بخشی از مجموعه پارامترهای خروجی در نظر گرفته شده‌اند. ولی در پژوهش حاضر، با استفاده از رویکرد خوشه‌بندی داده‌ها قبل از ورود به مرحله آموزش، به برخی از چالش‌های بسیار مهم مرتبط با یادگیری عمیق داده‌محور پرداخته شده‌است. بعد از خوشه‌بندی مناسب، با اعمال آزمون آماری KS (Kolmogorov-Smirnov: KS) نشان داده می‌شود که در داده‌هایی که با این رویکرد جداسازی شده‌اند، سه مجموعه دارای توزیع آماری یکسان هستند. با پیمایش کمی و کیفی نتایج وارون‌سازی داده‌های مگنتوتلوریک با درصدهای مختلف داده‌ای، نشان داده می‌شود که تعداد نمونه‌های مورد نیاز برای آموزش شبکه به شکل موثری، با حفظ دقت پیش‌بینی و یا حتی افزایش آن، کاهش می‌یابد.

۲ روش پژوهش

۲-۱ تولید داده‌های آموزشی

در روش مگنتوتلوریک از نوسان‌های میدان‌های الکترومغناطیسی طبیعی که بطور همزمان ثبت می‌شوند، برای کاوش ساختار ژئوالکتریکی زیرسطحی استفاده می‌شود. در حالت کلی تانسور امپدانس (Z) به عنوان مهمترین تابع تبدیل مگنتوتلوریک، رابطه بین مولفه‌های افقی میدان‌های الکتریکی (E) و مغناطیسی (H) را بیان می‌کند:

$$Z = \frac{E}{H} \qquad (۱)$$

در محیط یک‌بعدی، امپدانس الکتریکی یک کمیت عددی است و به سیستم مختصات بستگی ندارد؛ غالبا به شکل مقاومت ویژه ظاهری (ρ_a) و فاز (φ) بیان می‌شود:

$$\rho_a = \frac{1}{\omega\mu_0}|Z|^2 \quad , \quad \varphi = \tan^{-1}(\frac{ImZ}{ReZ}) \qquad (۲)$$

که μ_0 نفوذپذیری فضای آزاد و ω بسامد زاویه‌ای است.

برای تولید مثال‌های آموزشی از یک مدل ژئوالکتریکی که با حذف تغییرات جانبی از مدل به‌کاررفته توسط چن و همکاران (۲۰۱۲) حاصل شده‌است و شرایط یک میدان زمین‌گرمایی را شبیه‌سازی می‌کند، استفاده شده‌است. این مدل از پنج لایه شامل روباره، کلاهک رسی، مخزن، ناحیه گذار و سنگ بستر تشکیل شده است. آنها کران‌های اولیه برای پارامترهای مدل را بر اساس تجارب خود از میدان‌های زمین‌گرمایی انتخاب کرده‌اند و با استفاده از این کران‌ها می‌توان بازه‌هایی را برای تغییرات مقادیر مقاومت ویژه الکتریکی و عمق لایه‌ها تعیین کرد. مقادیر مقاومت ویژه الکتریکی به‌ترتیب از سطحی‌ترین به عمیق‌ترین لایه به‌طور تصادفی در بازه‌های [۶۰,۱۴۰]، [۱/۶,۲/۴]، [۲۰,۴۶۰]، [۶۴,۹۶] و [۳۰] اهم-متر توزیع شده‌اند. مقادیر ضخامت هم به‌طور تصادفی در بازه‌های [۴۵۰,۵۵۰]، [۱۷۰,۴۳۰]، [۱۷۷۰,۲۶۳۰] و [۱۰۷۰,۲۹۳۰] متر در نظر گرفته شده‌اند (جدول ۱). با استفاده از الگوریتم پیشرو MT پاسخ‌های این مدل ژئوالکتریکی یک‌بعدی به‌طور تحلیلی در بازه فرکانسی ۱۰۰-۰/۰۱ هرتز و در ۱۳ فرکانس که به‌طور یکنواخت در مقیاس لگاریتمی توزیع شده‌اند، محاسبه و در مجموع ۵۰۰۰۰۰ نمونه داده تولید شد.آماده‌سازی یا پیش‌پردازش داده‌ها شامل مراحلی همچون نرمال‌سازی و استانداردسازی به‌منظور مقیاس کردن مجدد متغیرهای ورودی و خروجی قبل از آموزش یک مدل شبکه است. متغیرهای ورودی با مقادیر خیلی پراکنده یا با واحدهای مختلف، در نهایت باعث می‌شوند مقادیر وزن تغییرات شدید داشته باشند و فرایند یادگیری ناپایدار شود. برای

سرعت بخشیدن به همگرایی شبکه، داده‌های ورودی (d_i) و داده‌های خروجی متناظر با آنها (m_j) به‌صورت زیر نرمال‌سازی می‌شوند تا در بازه $[۰ ,۱]$ واقع شوند:

$$\begin{cases} \hat{d}_i = \dfrac{\log(d_i)}{\log(d_{max})} \\ \hat{m}_j = \dfrac{\log(m_j)}{\log(m_{max})} \end{cases} \quad (۳)$$

که d_{max} و m_{max} به‌ترتیب معرف بیشینه مقادیر عددی داده ورودی و خروجی متناظر با آن هستند. گرچه روش‌های مختلفی برای نرمال‌سازی وجود دارد، این روش برای داده‌های MT مناسب‌تر به نظر می‌رسد. پس از اتمام مدل‌سازی شبکه، نتایج پیش‌بینی‌شده (m_j) توسط رابطه (۴) پساپردازش می‌شوند تا به بازه اولیه بازگردانده شوند و مقادیر خروجی مطلوب را ارائه دهند.

$$\hat{m}_j = anti\_\log(\log(m_{max}) \times m_j) \quad (۴)$$

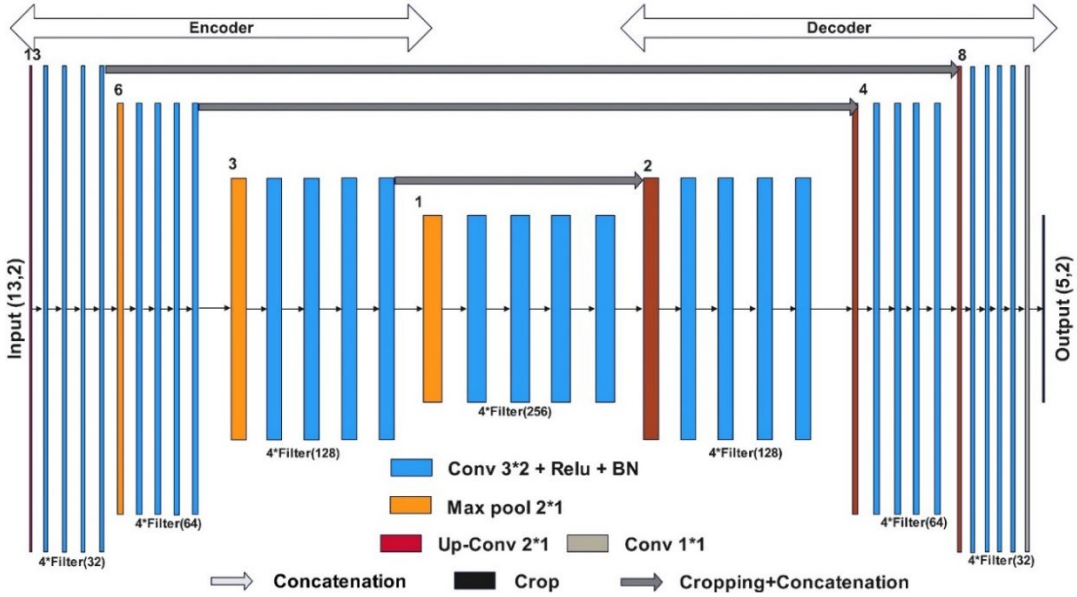
در مطالعات قبلی، به دلیل مشاهده عملکرد بهتر برای شبکه مورد نظر، فقط از مقاومت ویژه ظاهری به‌عنوان داده ورودی استفاده شده‌است (لیائو و همکاران، ۲۰۲۲؛ لیو و همکاران، ۲۰۲۱)؛ اما در این تحقیق مقادیر فاز هم به داده‌های ورودی اضافه شده‌اند و عملکرد شبکه طراحی شده را در انجام وارون‌سازی ارتقا داده‌اند.در این مرحله از یادگیری باناظر و طرحواره پس انتشار خطا استفاده می‌شود و بنابراین ورودی‌ها یا مشاهدات به همراه خروجی‌های مربوطه در قالب نمونه‌های آموزشی به شبکه داده می‌شوند.

جدول ۱. پارامترهای مورد استفاده برای تولید مدل ژئوالکتریکی یک‌بعدی از مخزن زمین‌گرمایی (برگرفته از مطالعه انجام شده توسط چن و همکاران، ۲۰۱۲).

لایه‌ها	بازه مقادیر مقاومت ویژه الکتریکی (اهم–متر)	بازه عمقی (متر)
لایه اول	[۶۰,۱۴۰]	[۴۵۰,۵۵۰]
لایه دوم	[۱/۶,۲/۴]	[۱۷۰,۴۳۰]
لایه سوم	[۲۰,۴۶۰]	[۱۷۷۰,۲۶۳۰]
لایه چهارم	[۶۴,۹۶]	[۱۰۷۰,۲۹۳۰]
لایه پنجم	[۳۰]	

۲-۲ معماری شبکه

شبکه‌های عصبی پیچشی (Convolutional Neural Network: CNN) در حال حاضر به‌طور گسترده مورد استفاده قرار می‌گیرند. تفاوت شبکه‌های CNN نسبت به شبکه‌های سنتی، استفاده از پیچش به جای ضرب ماتریسی است. دو اثر مشخص این کار، ایجاد اتصالات تنک و به اشتراک‌گذاری پارامتر است. یکی از پرکاربردترین شبکه‌های پیچشی، شبکه U-Net است.این معماری در ابتدا توسط رونبرگر و همکاران (۲۰۱۵) برای تقسیم‌بندی تصاویر پزشکی توسعه داده شد. اساس آن بخش‌های رمزنگار و رمزگشایی هستند که بین آنها پل ارتباطی برقرار شده و جریان اطلاعات را تسهیل می‌کند. در معماری U-Net به‌کاررفته در این پژوهش، از ۴ فیلتر با اندازه ۳۲، ۴ فیلتر با اندازه ۶۴، ۴ فیلتر با اندازه ۱۲۸ و ۴ فیلتر با اندازه ۲۵۶ استفاده شده‌است (شکل ۱). برای مسئله وارون 1-DMT ، داده‌های ورودی و خروجی ماتریس دوبعدی هستند؛ ماتریس خروجی به‌صورت داده‌های مقاومت‌ویژه الکتریکی و ضخامت و ماتریس ورودی، به‌صورت داده‌های مقاومت ویژه ظاهری و فاز تعریف شده‌اند. با وجود اینکه این معماری در مقایسه با سایر مدل‌های یادگیری عمیق با مجموعه داده‌های محدود نیز موثر است، همچنان چالش مربوط به جمع‌آوری حجم زیادی از داده‌ها و منابع سخت‌افزاری و زمان مورد نیاز برای انجام آن پابرجاست.



شکل ۱. معماری U-Net مورد استفاده.

۲-۳ آموزش شبکه یادگیری عمیق

توانایی نگاشت غیرخطی رویکردهای عمیق باعث می‌شود شبکه با تنظیم پارامترهای خود، قادر به ایجاد نگاشت بین داده‌های مقاومت ویژه الکتریکی ظاهری و فاز از یک طرف و مدل لایه‌ای مقاومت ویژه الکتریکی از طرف دیگر باشد. این پارامترها با به‌روزرسانی موثر، ویژگی‌های اصلی را از داده‌های ورودی فرا می‌گیرند. مقادیر اولیه پارامترها به شکل تصادفی انتخاب و در چرخه آموزش به‌روزرسانی می‌شوند. میزان کارایی روش‌های یادگیری ماشین به‌عنوان رویکردهای داده‌محور همچنین به فراپارامترهایی (hyperparameters) بستگی دارد که قبل از فرایند یادگیری انتخاب می‌شوند. آموزش شبکه نوعی فرایند بهینه‌سازی است و فراپارامترها بر اساس خطای اعتبارسنجی انتخاب می‌شوند. فراپارامترها ویژگی الگوریتم یادگیری هستند؛ در مقابل پارامترها ویژگی مدل هستند و بطور محض در حین آموزش، زمانی که الگوریتم سعی در یادگیری نگاشت بین ورودی و خروجی را دارد، تخمین زده می‌شوند. تعداد تکرار چرخه آموزش (Epoch)، اندازه دسته‌های نمونه‌گیری (Batch Size)، تعداد و اندازه فیلتر، نرخ یادگیری (Learning Rate)، الگوریتم بهینه‌سازی، تابع فعال‌ساز و تابع اتلاف از جمله فراپارامترهایی هستند که قبل از فرآیند یادگیری انتخاب و بعد از اتمام چرخه آموزش در صورت نیاز بر اساس منحنی خطای آموزش و اعتبارسنجی تغییر می‌کنند. در آموزش این شبکه با معماری U-Net، تعداد تکرار چرخه آموزش برابر با ۱۵، اندازه دسته‌های نمونه‌گیری برابر با ۳۲، تعداد و اندازه فیلتر برابر با ۴ فیلتر به اندازه های ۳۲، ۶۴، ۱۲۸و ۲۵۶، نرخ یادگیری برابر با ۰/۰۰۰۱، نوع تابع بهینه‌ساز از نوع Adam، تابع فعال ساز از نوع ReLU و تابع اتلاف از نوع میانگین مربعات خطا (MSE) در نظر گرفته شده است.

از تابع اتلاف برای محاسبه میزان خطا، از بهینه ساز Adam برای به روز رسانی وزن‌ها و از تابع فعال ساز ReLU برای انتقال وزن‌ها به لایه بعدی استفاده می‌شود. مجموعه پارامترهایی که در حین فرایند یادگیری بهینه می‌شوند، تابع نگاشت یا عملگر شبه‌وارونی را تشکیل می‌دهند که خروجی مدل یادگیری عمیق را با اعمال بر داده‌های ورودی تولید می‌کند (کیم و ناکاتا، ۲۰۱۸). در تابع بهینه‌ساز Adam برخلاف طرحواره کلاسیک کاهش

گرادیانی تصادفی (Stochastic Gradient Descent: SGD)، نرخ یادگیری ثابت نیست و به‌صورت تدریجی کم می‌شود. این رویه، واپاشی نرخ یادگیری نام دارد. یکی از مهم‌ترین مزیت‌های بهینه‌ساز Adam این است که نیاز کمتری به تنظیم دستی نرخ یادگیری دارد و در بسیاری از موارد، تنظیمات پیش‌فرض ($\alpha = ۰/۰۰۱$، $\beta_۱ = ۰/۹$ و $\beta_۲ = ۰/۹۹۹$) بهترین عملکرد را ارائه می‌دهند. بهینه‌ساز Adam به خوبی با داده‌های متفاوت و شرایط مختلف آموزشی سازگار است و توانایی آن در تطبیق نرخ یادگیری برای پارامترهای مختلف، به‌ویژه در مواجهه با داده‌های نامتعادل یا پراکنده بسیار ارزشمند است (کینگما و بار، ۲۰۱۵). مقدار اولیه نرخ یادگیری بر اساس سعی و خطا برابر با ۰/۰۰۰۱ انتخاب شد.

هنگام آموزش یک شبکه عمیق با یادگیری مبتنی بر گرادیان و طرحواره پس‌انتشار خطا (Backpropagation Error)، انباشت مشتقات کوچک منجر به مشکل محوشدگی گرادیان (Vanishing Gradient) می‌شود که به مدل ناپایدار و ناتوان در یادگیری موثر می‌انجامد. تابع فعال‌ساز ReLU، مشکل محوشدگی و انفجار گرادیان را که برای توابعی همچون سیگموید و تانژانت هذلولی رخ می‌دهد، ندارد. همچنین این تابع محدودیتی برای مقدار بیشینه ورودی ایجاد نمی‌کند. از آنجاییکه این تابع فقط تعدادی از گره‌ها را فعال می‌کند، از لحاظ محاسباتی از کارآیی بهتری نسبت به سایر توابع فعال‌ساز در شبکه‌های عمیق برخوردار است (نیر و همکاران، ۲۰۱۰).

یک ابزار تشخیصی و پرکاربرد در ارزیابی چرخه یادگیری در مسائل رگرسیون، پایش روند منحنی‌های خطای آموزش و اعتبارسنجی بعد از اتمام چرخه آموزش است. با محاسبه خطا پس از هر چرخه آموزش می‌توان نمودارهایی از عملکرد یادگیری بر مجموعه داده آموزشی و اعتبارسنجی ترسیم و از سه حالت بیش‌برازش (Overfit)، کم‌برازش (Underfit) و برازش مناسب (Goodfit) اطلاع حاصل کرد. همچنین می‌توان از این منحنی‌های خطا در تشخیص توزیع آماری یکسان داده‌های آموزش و اعتبارسنجی استفاده کرد (گودفلو و همکاران ، ۲۰۱۶).

۲-۴ منحنی‌های خطا در چرخه آموزش یادگیری عمیق

به‌طور کلی، منحنی خطا نموداری است که میزان خطا را برحسب تکرار در چرخه آموزش نشان می‌دهد. ترسیم منحنی خطا برای مجموعه داده آموزشی، ایده‌ای از چگونگی یادگیری مدل و ترسیم منحنی خطا برای مجموعه داده اعتبارسنجی، ایده‌ای از چگونگی تعمیم‌پذیری مدل ارائه می‌دهد. در ارزیابی چرخه آموزش، ترسیم همزمان این دو منحنی معمول است و از شکل و پویایی منحنی خطا می‌توان برای تشخیص رفتار یک مدل یادگیری عمیق استفاده کرد.

۲-۴-۱ منحنی‌های خطای کم‌برازش

کم برازش هنگامی اتفاق می‌افتد که مدل یادگیری عمیق به‌اندازه کافی پیچیده نباشد که بتواند رابطه‌ی میان ورودی و خروجی را یاد بگیرد و الگوی غالب در داده‌ها را تشخیص دهد؛ درنتیجه، در آموزش داده‌های آموزشی و در نتیجه اعتبارسنجی، عملکرد ضعیفی دارد. اگر مدلی نتواند به‌خوبی به داده‌های جدید تعمیم داده شود، نمی‌توان از آن برای طبقه‌بندی یا پیش‌بینی استفاده کرد. شکل منحنی خطا در مدل کم برازش معمولاً به دو صورت است:

۱- منحنی خطای آموزش بدون توجه به فرایند آموزش ثابت می ماند.

۲- منحنی خطای آموزش تا پایان چرخه آموزش همچنان کاهش می یابد.

چندین روش برای جلوگیری از کم‌برازش رخ‌داده پیشنهاد می‌شود که باعث افزایش پیچیدگی و تنوع مدل می‌شود و امکان آموزش موفقیت‌آمیز مدل فراهم می‌گردد.

۲-۴-۲ منحنی‌های خطای بیش‌برازش

برازش بیش از حد هنگامی اتفاق می‌افتد که مدل، ویژگی‌های داده‌های آموزشی را به‌جای یادگیری، حفظ کرده باشد؛ یعنی بیش از حد روی آن آموزش دیده باشد؛ درنتیجه، این مدل فقط در مجموعه‌ی داده‌های آموزشی مفید خواهد بود و نمی‌توان آموزش را به مجموعه‌ی داده‌های دیگر که هنوز آن‌ها را ندیده است تعمیم داد. بنابراین خطای تعمیم‌پذیری افزایش می‌یابد. این افزایش خطای تعمیم‌پذیری را می‌توان با عملکرد مدل در مجموعه داده اعتبارسنجی اندازه‌گیری کرد. شکل منحنی‌های خطای اعتبارسنجی در مدل بیش برازش تا یک نقطه کاهش و دوباره شروع به افزایش می‌یابد. نقطه عطف (Inflection point)، نقطه‌ای است که می‌توان آموزش را در آن متوقف کرد. زیرا تکرار آموزش پس از آن نقطه، برازش بیش از حد را نشان می‌دهد. روش‌های مختلفی برای جلوگیری از آن وجود دارد از جمله: تقویت داده‌ها (Data Augmentation)، منظم سازی L2 (Regularization L2)، حذف تصادفی (Drop Out)، توقف زودهنگام (Early Stopping) و ذخیره بهترین وزن‌ها. در اینجا از منظم سازی L2 با مقدار ۰/۰۰۰۱ و ذخیره بهترین وزن‌ها استفاده شده است.

۲-۴-۳ منحنی‌های یادگیری برازش مناسب

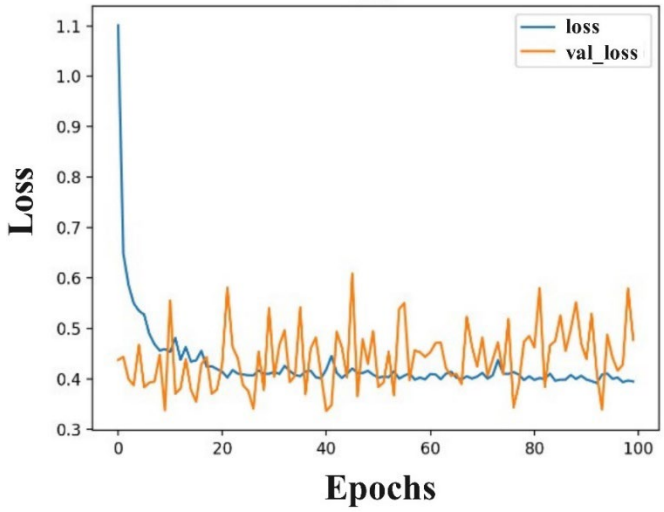
در برازش مناسب، فاصله بین خطای یادگیری آموزش و اعتبارسنجی کمینه بوده و هر دو منحنی روند نزولی دارند و در واقع برازش مناسب، هدف الگوریتم یادگیری است. بهترین انتخاب در تعداد تکرار چرخه آموزش، زمانی است که روند کاهشی منحنی‌های خطای آموزش و اعتبارسنجی ناملموس باشد و فاصله بین آنها به کمترین مقدار خود برسد.

۲-۴-۴ تشخیص توزیع یکسان داده‌های آموزش و اعتبارسنجی

از منحنی‌های خطا می‌توان در تشخیص توزیع یکسان داده‌های آموزش و اعتبارسنجی استفاده کرد. در صورتی‌که ویژگی‌های آماری دو مجموعه داده آموزش و اعتبارسنجی یکسان نباشند برازش مناسب در چرخه آموزش اتفاق نمی‌افتد. اگر تعداد نمونه‌ها در یک مجموعه داده نسبت به مجموعه داده دیگر بسیار کم باشد ویژگی‌های آماری دو مجموعه یکسان نخواهد بود. معمولاً در جداسازی داده‌ها در یادگیری عمیق، سهم بیشتر به داده‌های آموزش اختصاص دارد و تعداد نمونه‌ها در مجموعه داده اعتبارسنجی کم است. در این حالت اگر ویژگی‌های آماری مجموعه داده اعتبارسنجی در مقایسه با مجموعه داده آموزشی یکسان نباشد، مجموعه داده اعتبارسنجی اطلاعات کافی به‌منظور ارزیابی توانایی مدل برای تعمیم‌پذیری را ارائه نمی‌دهد. در این حالت روند منحنی خطای آموزش، نزولی، اما روند منحنی خطای اعتبارسنجی نامنظم خواهد بود (شکل۲).

جدول ۲. تعداد داده‌ها در سه مجموعه آموزش، اعتبارسنجی و آزمایش با رهیافت تصادفی.

نسبت جداسازی تصادفی	تعداد داده آموزش	تعداد داده اعتبارسنجی	تعداد داده آزمایش
۹۸/۱/۱	۴۹۰۰۰۰	۵۰۰۰	۵۰۰۰
۹۰/۵/۵	۴۵۰۰۰۰	۲۵۰۰۰	۲۵۰۰۰
۷۰/۱۵/۱۵	۳۵۰۰۰۰	۷۵۰۰۰	۷۵۰۰۰

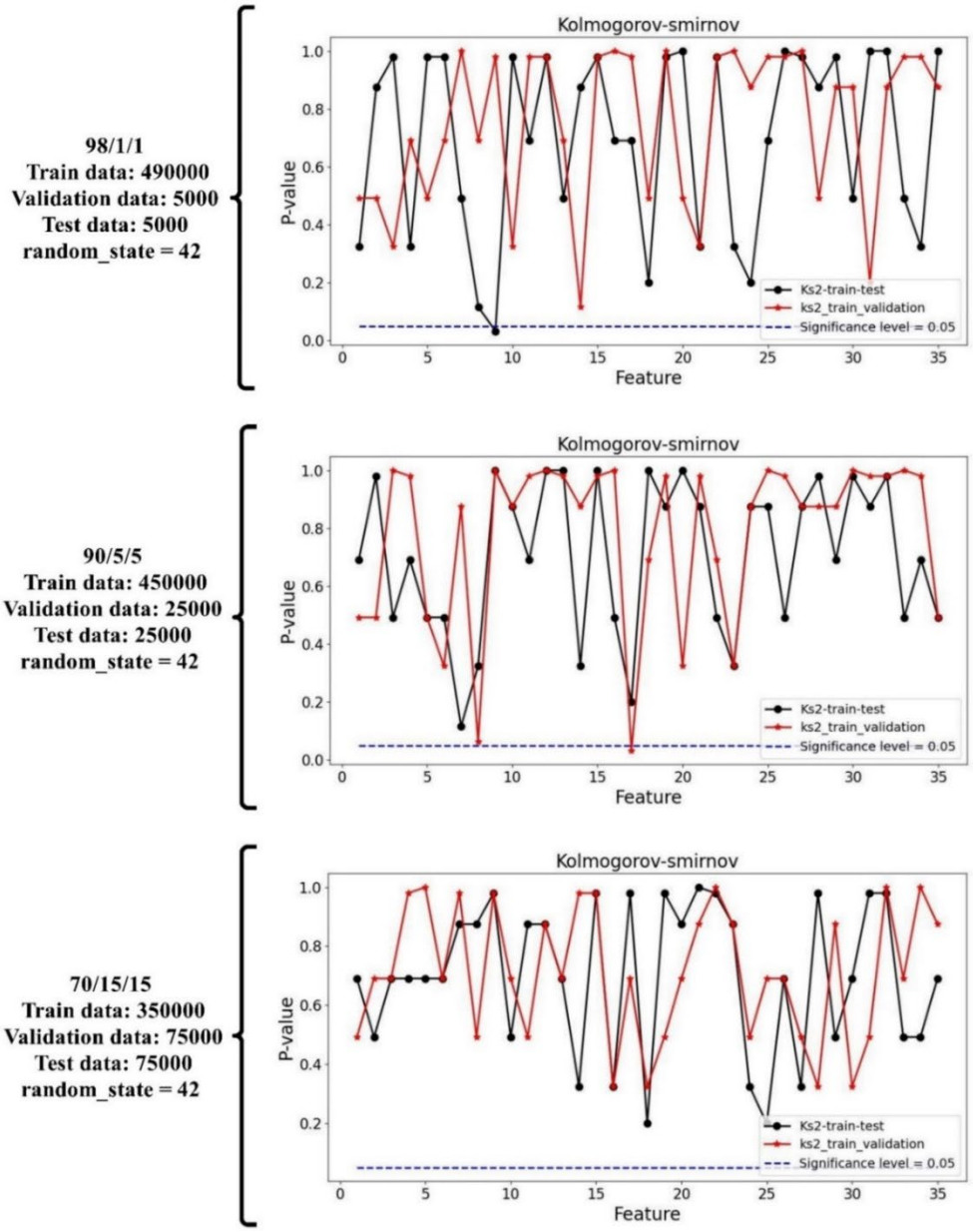


شکل ۲. رفتار نامنظم در منحنی خطای اعتبارسنجی به دلیل یکسان نبودن توزیع آماری داده‌های اعتبارسنجی و آموزش.

۳ خوشه‌بندی داده‌ها

یکی از مسائلی که در آموزش شبکه با آن مواجه هستیم، انتخاب نسبت جداسازی داده‌های آموزشی، اعتبارسنجی و آزمایش است. بر اساس مطالعات مرتبط با یادگیری عمیق، اگر تعداد داده‌ها زیاد باشد معمولا نسبت جداسازی به‌صورت ۹۸/۱/۱ و اگر تعداد داده‌ها کم باشد، به شکل ۷۰/۱۵/۱۵ در نظر گرفته می‌شود (اولانی مورینا، ۲۰۲۲)؛ ولی هیچ معیار مشخصی برای نحوه توزیع داده‌ها در سه مجموعه وجود ندارد و عموما به شکل تصادفی انجام می‌شود. حالت ایده‌آل این است که توزیع آماری سه مجموعه یکسان باشد؛ در غیر اینصورت منحنی خطای اعتبارسنجی رفتار نامنظمی را نشان می‌دهد.

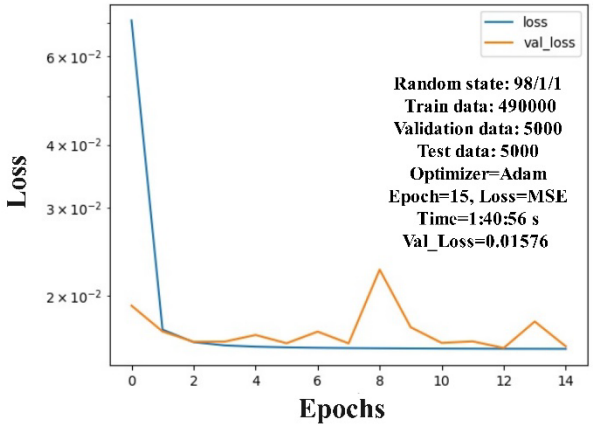
دو راهکار برای یکسان‌سازی توزیع آماری سه مجموعه داده وجود دارد: ۱- تعداد داده ها زیاد و نسبت جداسازی کم شود. ۲- با استفاده از روش‌های خوشه‌بندی توزیع آماری آنها یکسان شود. ابتدا این موضوع بررسی می‌شود که آیا با نسبت‌های مختلف امکان دستیابی به توزیع آماری یکسان برای سه مجموعه داده وجوددارد. آزمون KS یکی از آماره‌هایی است که برای بررسی یکسان بودن توزیع آماری دو مجموعه به‌کار می‌رود. در این بررسی فرض صفر مبتنی بر اینکه توزیع داده‌ها یکنواخت است، در سطح خطای ۰/۰۵ که سطح معناداری (significance level) نام دارد، آزمایش می‌شود. اگر آماره آزمون یا همان مقدار P-value) P) بزرگتر از یا مساوی با سطح خطای ۰/۰۵ باشد، در این‌صورت دلیلی برای رد فرض صفر وجود ندارد و توزیع داده‌ها یکنواخت است (فرجی، ۱۳۸۵). مقدار P در واقع اختلاف بین تابع توزیع تجربی هریک از ویژگی‌های دو مجموعه است و برای اینکه توزیع آماری دو مجموعه داده یکسان باشد، باید مقدار این آماره برای همه ویژگی‌ها بزرگتر از سطح معناداری باشد. هر نمونه داده تولید شده در بخش ۲-۱ با ۳۵ ویژگی (۲۶ ویژگی مرتبط با پاسخ‌های مدل‌سازی پیشرو شامل ۱۳ مقدار مقاومت ویژه ظاهری و ۱۳ مقدار فاز به‌اضافه ۹ ویژگی مرتبط با پارامترهای فیزیکی و هندسی مدل زمین‌گرمایی انتخاب شده) معرفی می‌شود. جدول ۲ تعداد داده‌ها در سه مجموعه داده آموزش، اعتبارسنجی و آزمایش در نسبت‌های مختلف جداسازی تصادفی ۹۸/۱/۱، ۹۰/۵/۵ و ۷۰/۱۵/۱۵ و شکل ۳ نتیجه اعمال آزمون KS برای ویژگی‌های مختلف و نسبت‌های متفاوت جداسازی تصادفی داده‌ها را نشان می‌دهد.



شکل ۳. نتیجه آزمون KS برای ارزیابی یکسان بودن توزیع آماری سه مجموعه داده آموزش، اعتبارسنجی و آزمایش با نسبت‌های مختلف جداسازی تصادفی ۹۸/۱/۱، ۹۰/۵/۵ و ۷۰/۱۵/۱۵ برای مجموعه داده ۵۰۰۰۰۰ تایی.

همان‌طور که در شکل ۳ ملاحظه می‌شود در نسبت‌های ۹۸/۱/۱ و ۹۰/۵/۵ آزمون KS پایین‌تر از سطح معنی‌دار ۰/۰۵ قرار دارد و بیانگر این است که توزیع آماری بین داده‌های آموزش و آزمایش در نسبت ۹۸/۱/۱ و بین داده‌های آموزش و اعتبارسنجی در نسبت ۹۰/۵/۵، یکسان نیست. با کم شدن نسبت جداسازی داده‌ها (۷۰/۱۵/۱۵) و افزایش تعداد داده‌های اعتبارسنجی، اساساً توزیع آماری داده‌ها یکسان می‌شود. البته شایان ذکر است جداسازی داده‌ها به صورت تصادفی در برنامه نویسی، ارتباط مستقیم با عدد حالت تصادفی (Random State) دارد و با تغییر عدد حالت تصادفی ممکن است در نسبت‌های ذکر شده، در توزیع آماری داده‌ها در سه مجموعه داده تغییر ایجاد

شود. شکل (۴) منحنی‌های خطای آموزش و اعتبارسنجی با معماری U-Net در جداسازی تصادفی با نسبت ۹۸/۱/۱ را نشان می‌دهد. در معماری U-Net از الگوریتم Adam و تنظیمات فراپارامترهای گفته‌شده در بخش ۲-۳ استفاده شده‌است. همانگونه که مشاهده می‌شود، به دلیل یکسان نبودن توزیع آماری دو مجموعه داده آموزش و اعتبارسنجی، روند منحنی خطای آموزش، نزولی، اما روند منحنی خطای اعتبارسنجی، نامنظم و دارای نوسان است و مدل تعمیم‌پذیری مناسبی را ارائه نمی‌دهد.



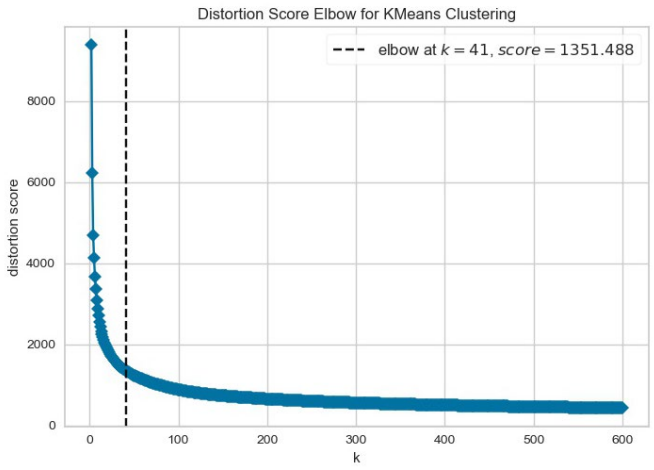
شکل ۴. رفتار نامنظم در منحنی خطای اعتبارسنجی به دلیل یکسان نبودن توزیع آماری داده‌های اعتبارسنجی و آموزش در نسبت تصادفی ۹۸/۱/۱.

رویکرد این مطالعه، اعمال روش‌های خوشه‌بندی بر داده‌ها و اختصاص درصد مشخصی از هر خوشه به سه مجموعه داده است. با این رویکرد، تعداد داده‌های مورد نیاز و زمان آموزش، کاهش و دقت پیش‌بینی افزایش می‌یابد. در ابتدا ۵۰۰۰۰۰ نمونه تولید شد (بخش ۲-۱) و هر نمونه داده با ۳۵ ویژگی، ۲۶ ویژگی مرتبط با پاسخ مدل‌سازی پیشرو یا ورودی و ۹ ویژگی مرتبط با پارامترهای مدل یا خروجی، معرفی می‌شود. الگوریتم K-means به‌عنوان یکی از پرکاربردترین الگوریتم‌های خوشه‌بندی بر داده‌های MT تولید شده اعمال و از نتایج خوشه‌بندی برای ایجاد و انتخاب مجموعه داده نمونه آموزشی استفاده شد. K-means یک الگوریتم مبتنی بر تکرار است که مجموعه داده را به زیرگروه‌ها یا خوشه‌های از قبل تعریف‌شده متمایز بدون همپوشانی تقسیم می‌کند که در آن هر نقطه داده فقط به یک گروه تعلق دارد. نقاط داده درون خوشه‌ای تا حد ممکن شبیه هم هستند و در عین حال خوشه‌ها تا حد ممکن متفاوت نگه داشته می‌شوند. نقاط داده طوری به یک خوشه اختصاص پیدا می‌کنند که مجموع مجذور فاصله بین نقاط داده و مرکزوار خوشه (میانگین حسابی تمام نقاط داده متعلق به آن خوشه)، به‌عنوان معیار تشابه کمینه باشد. نحوه پیاده‌سازی الگوریتم به این ترتیب است: تعداد خوشه‌ها K مشخص می‌شود؛ با انتخاب تصادفی K نقطه داده برای مرکزوارها بدون جایگزینی، مرکزوارها در حالت اولیه قرار می‌گیرند؛ تا زمانی که بعد از آن هیچ تغییری در مرکزوارها ایجاد نشود، یعنی تخصیص نقاط داده به خوشه‌ها تغییر نکند، تکرارها ادامه پیدا می‌کنند؛ مجموع مجذور فاصله بین نقاط داده و همه مرکزوارها محاسبه می‌شوند؛ هر نقطه داده به نزدیک‌ترین خوشه یا مرکزوار اختصاص داده می‌شود؛ با محاسبه میانگین تمام نقاط داده‌ای که به هر خوشه تعلق دارند، مرکزوارها برای خوشه‌ها محاسبه می‌شوند (مک‌کوین، ۱۹۶۷). منحنی زانو یک روش گرافیکی برای یافتن مقدار بهینه K است. در این نمودار

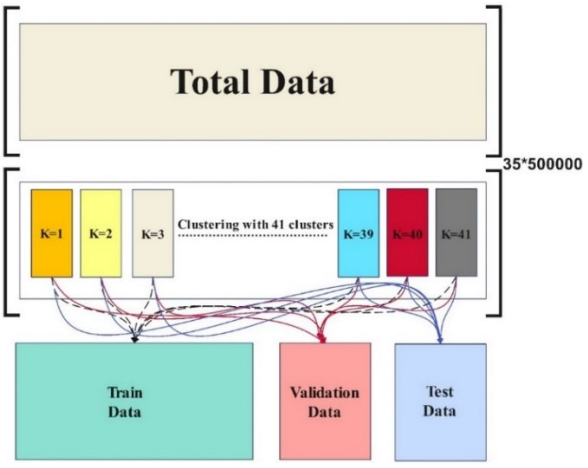
مجموع مربعات فواصل درون خوشه‌ای بر حسب مقادیر مختلف K نمایش داده می‌شود (کودیناریا و مکوانا، ۲۰۱۳). در شکل ۵ منحنی زانو برای مقادیر K بین ۲ تا ۶۰۰ ترسیم شده‌است و مقدار بهینه K مربوط به نقطه‌ای است که در آن یک زانو شکل می‌گیرد. طبق شکل تعداد بهینه خوشه‌ها برابر با ۴۱ است. بنابراین داده‌ها در قالب ۴۱ دسته، خوشه‌بندی می‌شوند و از هر خوشه با درصدهای متفاوتی داده‌ها به سه مجموعه آموزش، اعتبارسنجی و آزمایش اختصاص داده می‌شوند (شکل ۶). هدف، دستیابی به بهترین نتیجه با کمترین تعداد داده است. جدول ۳ تعداد داده‌ها را در سه مجموعه آموزش، اعتبارسنجی و آموزش با رویکرد خوشه‌بندی نشان می‌دهد.

جدول ۳. تعداد داده‌ها در سه مجموعه آموزش، اعتبارسنجی و آزمایش با رویکرد خوشه‌بندی.

درصدهای متفاوت از نسبت جداسازی ۹۸/۱/۱	تعداد داده آموزش	تعداد داده اعتبارسنجی	تعداد داده آزمایش
۹۸/۱/۱	۴۹۸۹۸۲	۴۹۸۰	۴۹۸۰
۱۰٪(۹۸/۱/۱)	۴۸۹۷۹	۴۸۲	۴۸۲
۲۰٪(۹۸/۱/۱)	۹۷۹۷۸	۹۸۱	۹۸۱
۳۰٪(۹۸/۱/۱)	۱۴۶۹۷۹	۱۴۷۸	۱۴۷۸
۴۰٪(۹۸/۱/۱)	۱۹۵۹۸۲	۱۹۸۱	۱۹۸۱
۵۰٪(۹۸/۱/۱)	۲۴۴۹۸۱	۲۴۷۹	۲۴۷۹

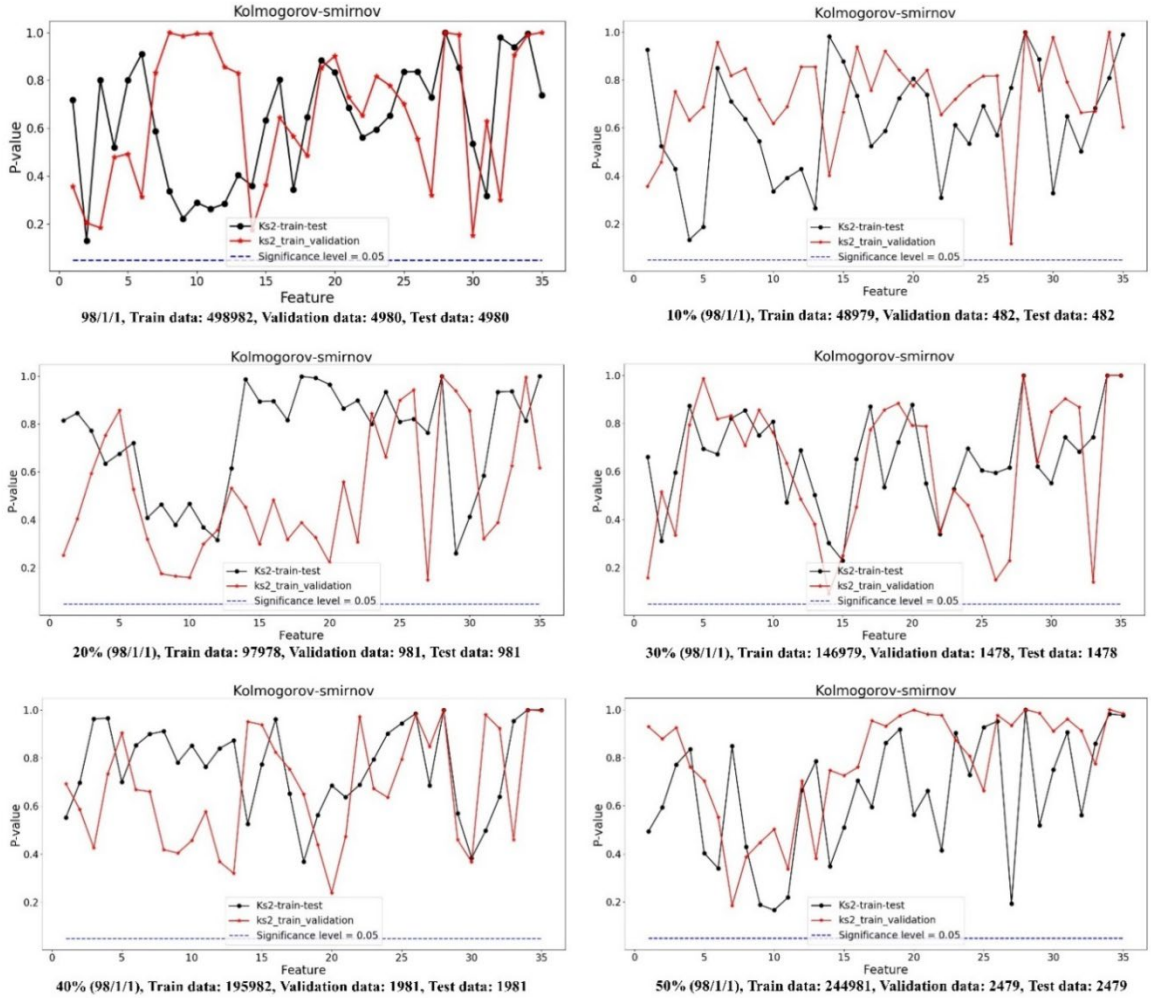


شکل ۵. منحنی مجموع مربعات فواصل درون خوشه‌ای بر حسب مقادیر مختلف K.



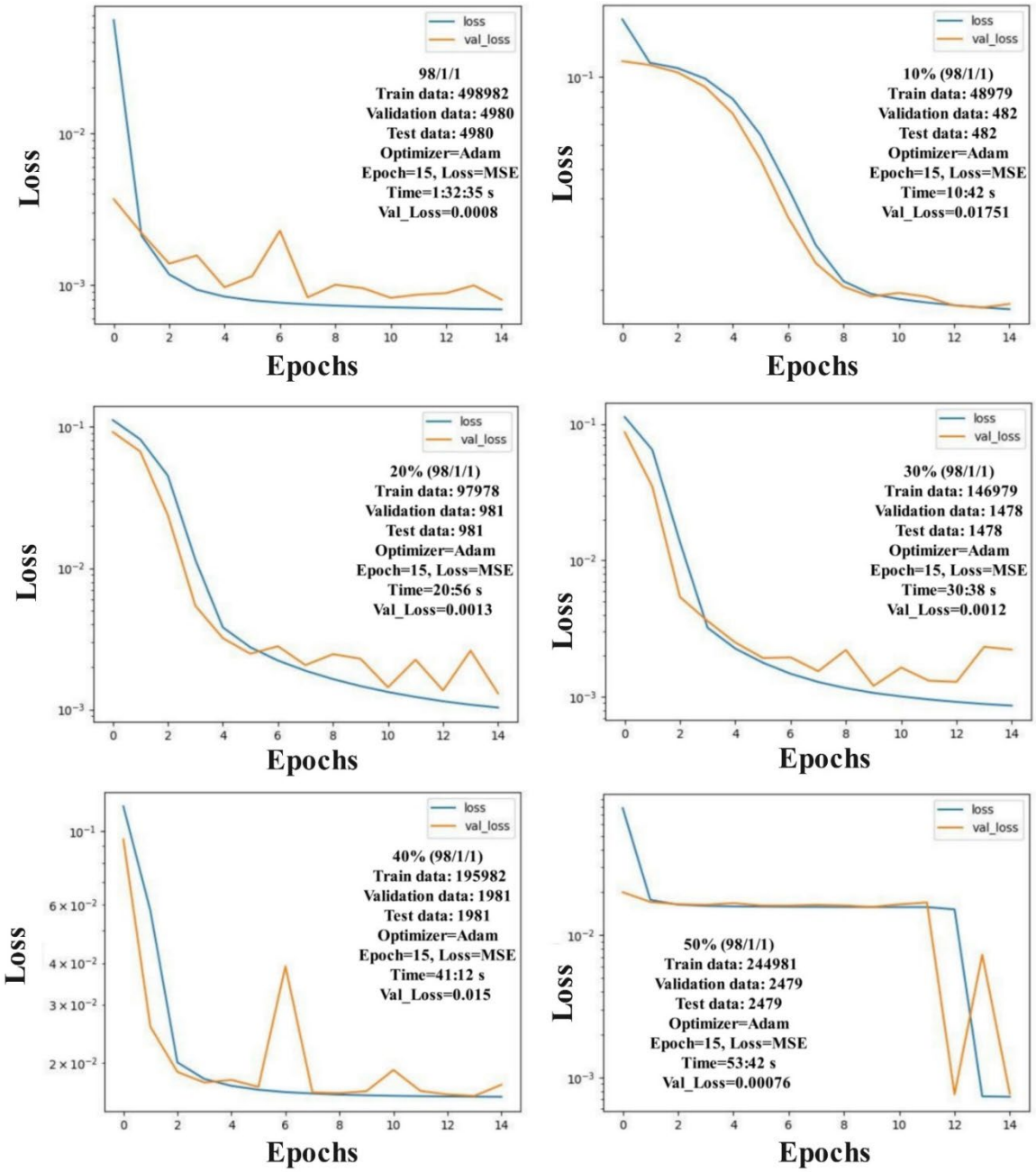
شکل ۶. پیاده‌سازی الگوریتم Kmeans برای خوشه‌بندی داده‌های آموزشی MT.

قبل از آموزش شبکه با داده‌های خوشه‌بندی‌شده، توزیع آماری سه مجموعه داده در حالت‌های متفاوت مقایسه می‌شوند. به این منظور از آزمون KS بهره گرفته می‌شود. نتایج در شکل ۷ نشان داده شده‌اند. چنانکه مشاهده می‌شود در همه حالت‌ها آماره آزمون KS بزرگتر از سطح معنادار ۰/۰۵ است و معرف یکسان بودن توزیع داده‌ها با نسبت‌های داده‌ای متفاوت است.



شکل ۷. نتایج آزمون KS برحسب ویژگی برای درصدهای متفاوت از کل داده‌ها در نسبت ۹۸/۱/۱.

در مرحله بعد، داده‌های حاصل از خوشه‌بندی با درصدهای متفاوت از مقدار کل و نسبت جداسازی ۹۸/۱/۱ با معماری U-Net و الگوریتم Adam و تنظیمات گفته‌شده در بخش ۲-۳ آموزش داده می‌شوند و نتایج با حالتی که کل داده‌ها با نسبت جداسازی ۹۸/۱/۱ وارد شده‌اند، مقایسه می‌شوند. منحنی‌های خطای آموزش و اعتبارسنجی که در یادگیری عمیق پایش می‌شوند، در شکل ۸ نمایش داده‌شده‌اند. چنانکه مشاهده می‌شود در حالتی که ۵۰ درصد ازداده‌های کل به‌کار گرفته شده‌اند، کمترین خطای یادگیری محقق می‌شود.



شکل ۸. منحنی‌های خطای آموزش و اعتبارسنجی برای درصدهای متفاوت از کل داده‌ها.

همانطور که در شکل ۸ ملاحظه می‌شود، در نسبت‌های ۹۸/۱/۱، ۱۰٪(۹۸/۱/۱)، ۲۰٪(۹۸/۱/۱)، ۵۰٪(۹۸/۱/۱) برازش خوب و در نسبت‌های ۳۰٪(۹۸/۱/۱) و ۴۰٪(۹۸/۱/۱) برازش بیش از حد اتفاق افتاده است که با فراخوانی تابع ذخیره بهترین وزن‌ها در پایش منحنی خطای اعتبارسنجی، در تکراری که خطای اعتبارسنجی کمترین مقدار را دارد ماتریس وزن نورون‌ها به عنوان بهترین مدل ذخیره می‌شود. نتایج وارون‌سازی داده‌های حاصل از خوشه‌بندی توسط معماری U-Net در جدول ۴ نشان می‌دهد که کمترین خطای آموزش در نسبت‌های ۹۸/۱/۱،

(۹۸/۱/۱)٪۳۰ و (۹۸/۱/۱)٪۵۰ اتفاق افتاده‌است. عملکرد یادگیری عمیق در نسبت‌های ۹۸/۱/۱ و (۹۸/۱/۱)٪۵۰ به صورت برازش خوب و در نسبت (۹۸/۱/۱)٪۳۰ به صورت بیش برازش است. در نسبت (۹۸/۱/۱)٪۳۰ ذخیره بهترین وزن‌ها با کمینه خطای اعتبار سنجی ۰/۰۰۱۲ در تکرار ۱۳ بوده‌است. در نسبت‌های ۹۸/۱/۱ و (۹۸/۱/۱)٪۵۰ کمینه خطای اعتبارسنجی نزدیک به هم هستند؛ بنابراین در آموزش شبکه در نسبت داده‌ای (۹۸/۱/۱)٪۵۰ با زمان آموزش کمتر و با تعداد داده کمتر، نتایج مشابه با نسبت داده‌ای ۹۸/۱/۱ حاصل خواهد شد.

جدول ۴. نتایج وارون‌سازی داده‌های حاصل از خوشه‌بندی توسط معماری U-Net.

درصدهای متفاوت از نسبت جداسازی ۹۸/۱/۱	عملکرد یادگیری	کمینه خطای آموزش	کمینه خطای اعتبارسنجی	شماره تکرار در ذخیره بهترین مدل	مدت زمان آموزش
۹۸/۱/۱	برازش خوب	۰/۰۰۰۷	۰/۰۰۰۸	۱۵	۱ : ۳۲ : ۳۵ s
۱۰٪(۹۸/۱/۱)	برازش خوب	۰/۰۱۵	۰/۰۱۷۵۱	۱۵	۱۰ : ۴۲ s
۲۰٪(۹۸/۱/۱)	برازش خوب	۰/۰۰۱	۰/۰۰۱۳	۱۵	۲۰ : ۵۶ s
۳۰٪(۹۸/۱/۱)	بیش‌برازش	۰/۰۰۰۸	۰/۰۰۱۲	۱۳	۳۰ : ۳۸ s
۴۰٪(۹۸/۱/۱)	بیش‌برازش	۰/۰۱۴	۰/۰۱۵	۱۴	۴۱ : ۱۲ s
۵۰٪(۹۸/۱/۱)	برازش خوب	۰/۰۰۰۷	۰/۰۰۰۷۶	۱۵	۵۳ : ۴۲ s

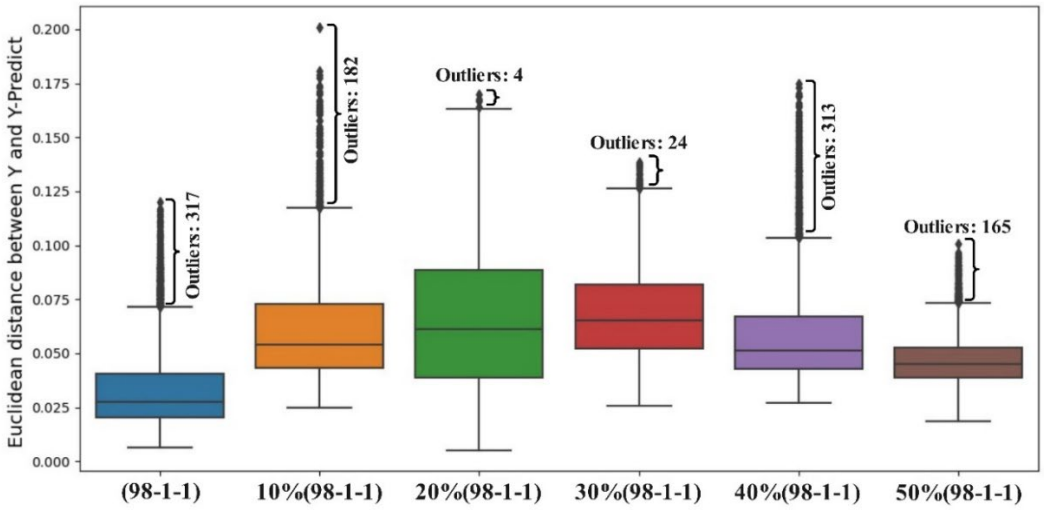
۴ بحث و ارزیابی نتایج

۴-۱ بررسی کیفی نتایج وارون‌سازی

برای بررسی کیفی نتایج وارون‌سازی، از یک مجموعه داده آزمایشی یکسان (با ۴۹۸۰ نمونه داده) برای پیش بینی نتایج استفاده شد. از آنجایی که پیش‌بینی نتایج، نوعی رگرسیون دو متغیره است (مقاومت ویژه هر لایه همراه با ضخامت آن) به دلیل ماهیت متفاوت دو کمیت مقاومت ویژه و ضخامت، ابتدا مقادیر داده‌های پیش‌بینی و واقعی نرمال‌سازی شده و سپس از فاصله اقلیدوسی برای بررسی اختلاف مقادیر پیش‌بینی شده و واقعی استفاده می‌شود. نمودار جعبه‌ای، یک روش استاندارد برای نمایش توزیع یک مجموعه داده بر اساس پنج ویژگی عددی کمینه، چارک اول، میانه، چارک سوم و بیشینه آن نقاط داده است. بنابراین ابزار مفیدی برای نشان دادن مقادیر میانگین، میزان تقارن، مقادیر پرت و پراکندگی مجموعه داده‌ها محسوب می‌شود.فاصله اقلیدسی در قالب نمودار جعبه‌ای برای شش نسبت مختلف در شکل ۹ ترسیم شده است. میانه و دامنه میان‌چارکی نمودار جعبه‌ای با نسبت داده‌ای (۹۸/۱/۱)٪۵۰، بیشترین نزدیکی را با نسبت داده‌ای ۹۸/۱/۱ دارد ؛ ضمناً بیشترین تقارن هم در همین حالت مشاهده می‌شود، هرچند که کمترین تعداد داده پرت مربوط به نسبت (۹۸/۱/۱)٪۲۰ است.

۴-۲ بررسی کمی نتایج وارون‌سازی

ضریب بهره‌وری ناش‌ساتکلیف (Nash-Sutcliffe efficiency: NSE) در سال ۱۹۷۰ به منظور ارزیابی کمی درجه انطباق سری‌های زمانی مشاهده و پیش‌بینی‌شده معرفی شد (ناش و ساتکلیف، ۱۹۷۰). از این معیار در اینجا برای ارزیابی میزان برازش بین منحنی‌های مقاومت ویژه ظاهری و فاز پیش‌بینی‌شده و واقعی استفاده شده‌است. ضریب NSE نشان می‌دهد که خط رگرسیون بین داده‌های مشاهده و پیش‌بینی‌شده تا چه حد به خط رگرسیون با شیب



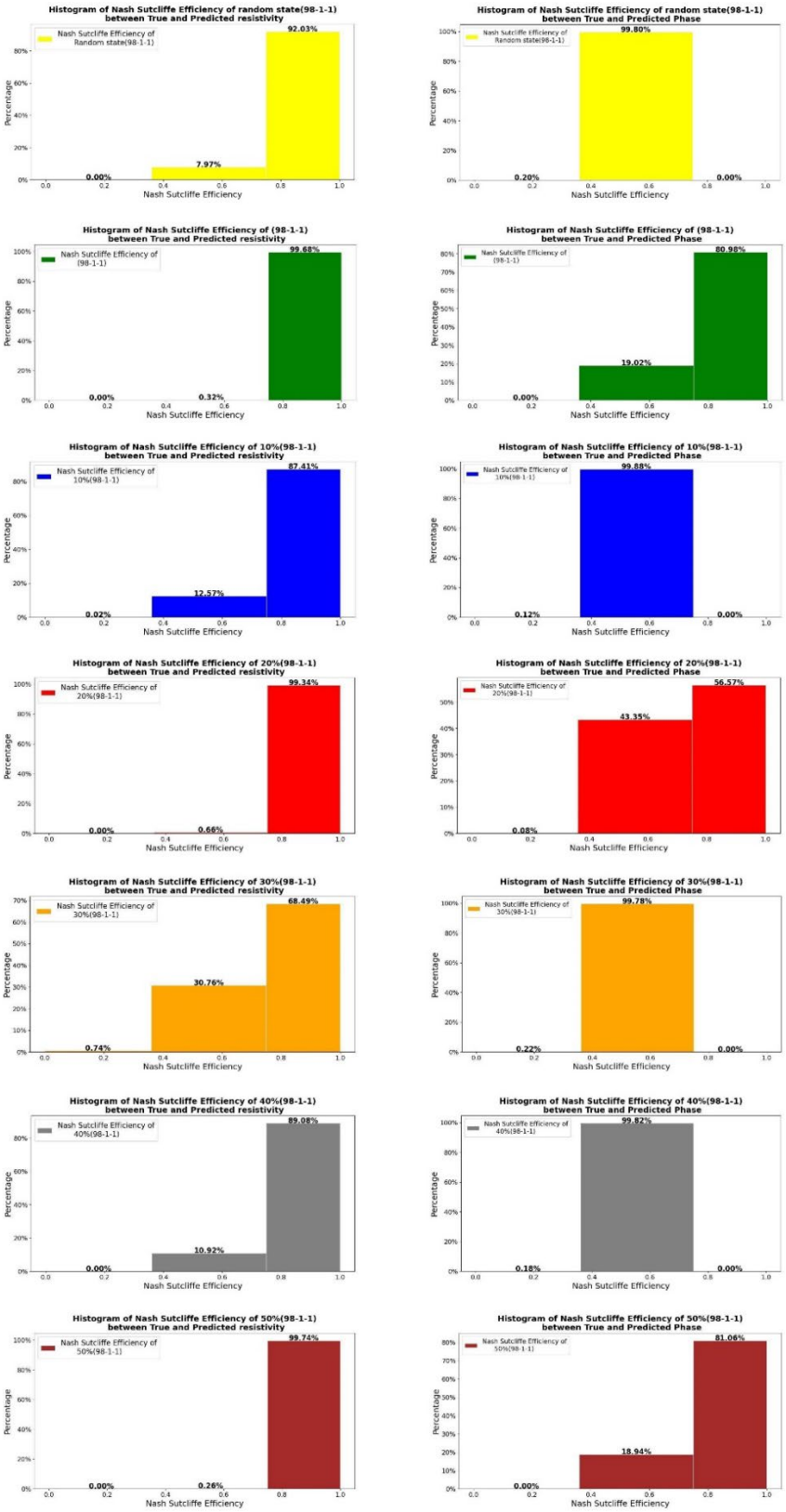
شکل ۹. نمودار جعبه‌ای فواصل اقلیدسی بین مقادیر واقعی و پیش‌بینی شده توسط شبکه عمیق.

یک نزدیک است. این ضریب بر اساس رابطه ۵ محاسبه می‌شود:

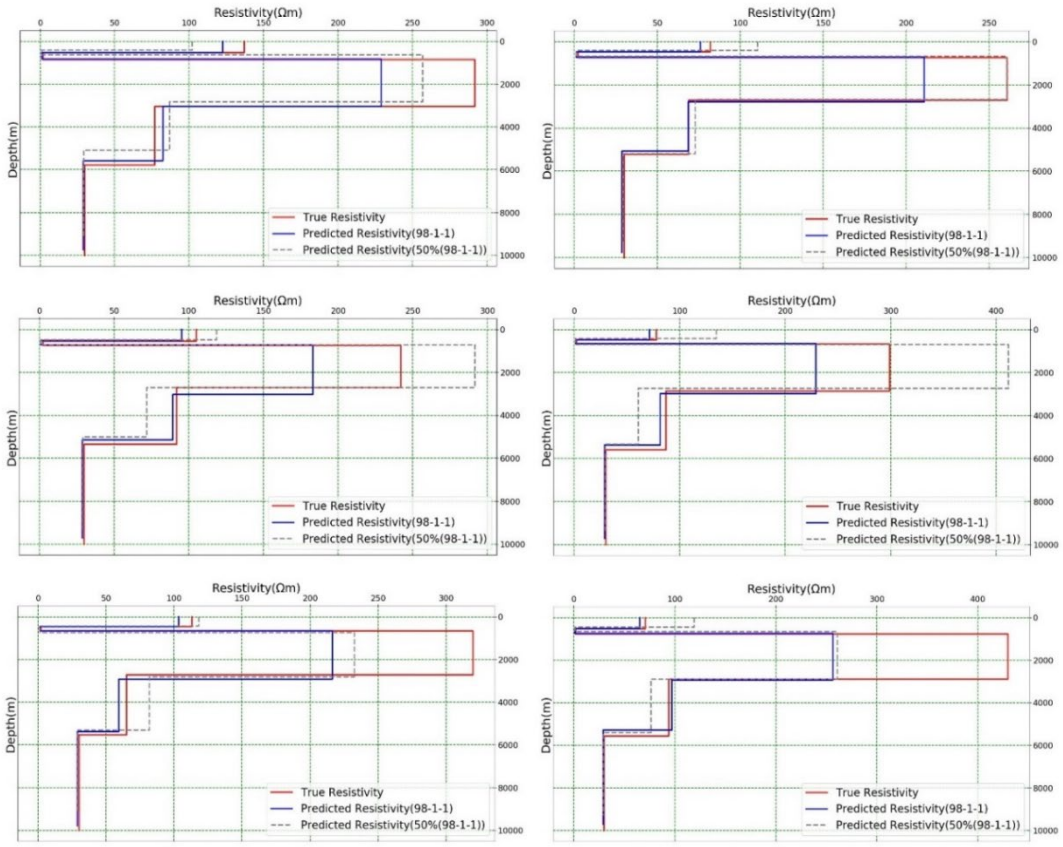
$$NSE_{\rho_a} = 1 - \left[\frac{\sum_i \left(\rho_{a_i}^{obs} - \rho_{a_i}^{pre}\right)^2}{\sum_i \left(\rho_{a_i}^{obs} - \rho^{mean}\right)^2}\right] \quad ; \quad NSE_{ph} \quad (۵)$$
$$= 1 - \left[\frac{\sum_i (ph_i^{obs} - ph_i^{pre})^2}{\sum_i (ph_i^{obs} - ph^{mean})^2}\right]$$

که در آن $\rho_{a_i}^{obs}$ و ph_i^{obs} مقاومت ویژه ظاهری و فاز مشاهده‌شده، $\rho_{a_i}^{pre}$ و ph_i^{pre} مقاومت ویژه ظاهری و فاز پیش‌بینی‌شده و ρ^{mean} و ph^{mean} میانگین مقاومت ویژه ظاهری و فاز مشاهده‌شده هستند. در دسته بندی ضریب ناش ساتکلیف، مقادیر کمتر از ۰/۳۶، ضعیف، مقادیر بین ۰/۳۶ تا ۰/۷۵ به عنوان خوب و مقادیر بیشتر از ۰/۷۵ عالی در نظر گرفته می‌شوند. برای ارزیابی کمی نتایج وارون‌سازی، ضریب NSE منحنی‌های مقاومت ویژه ظاهری و فازی که با استفاده از مدل‌سازی پیشرو از روی نتایج پیش‌بینی شبکه به‌دست آمده‌اند و داده‌های واقعی برای نسبت‌های جداسازی مختلف محاسبه و درصد فراوانی آنها رسم می‌شود. مقادیر NSE از بی‌نهایت منفی تا یک متغیر هستند؛ مقادیر بزرگ‌تر نشان‌دهنده عملکرد بهتر و مقادیر کوچک یا منفی معرف عملکرد ضعیف‌تر مدل هستند (واندرکلن و همکاران، ۲۰۱۸).

شکل (۱۰) هیستوگرام درصد فراوانی ضریب NSE برای منحنی‌های مقاومت ویژه ظاهری و فاز در جداسازی تصادفی و نسبت‌های مختلف داده‌ای در حالت خوشه‌بندی را نشان می‌دهد. در هیستوگرام منحنی‌های مقاومت ویژه ظاهری در جداسازی تصادفی (۹۸/۱/۱)، ۹۳/۰۳٪ نمونه‌ها دارای ضرائب بزرگتر از ۰/۷۵ و ۷/۹۳٪ نمونه‌ها دارای ضرائب بین ۰/۳۶ و ۰/۷۵ هستند؛ در هیستوگرام منحنی‌های فاز در همین حالت، ۹۹/۸٪ دارای ضرائب بین ۰/۳۶ و ۰/۷۵ و سایر نمونه‌ها دارای مقادیر کوچکتر از ۰/۷۵ هستند. در حالی‌که در هیستوگرام منحنی‌های مقاومت ویژه ظاهری در حالت خوشه‌بندی و در نسبت داده‌ای ۹۸/۱/۱، ۹۹/۶۸٪ نمونه‌ها ضرائب بزرگتر از ۰/۷۵ و در هیستوگرام منحنی‌های فاز در همین حالت، ۱۹/۰۲٪ نمونه‌ها ضرائب بین ۰/۳۶ و ۰/۷۵ و ۸۰/۹۹٪ نمونه‌ها ضرائب بزرگتر از ۰/۷۵ دارند. بنابراین نتایج وارون‌سازی با استفاده از داده‌های خوشه‌بندی‌شده، بسیار بهتر از نتایج وارون‌سازی با استفاده از داده‌هایی است که به‌شکل تصادفی جداسازی شده‌اند. با توجه به این شکل، از بین دیگر نسبت‌ها، تنها نسبتی که



شکل ۱۰. هیستوگرام درصد فراوانی ضرایب NSE برای منحنی‌های مقاومت ویژه ظاهری و فاز در حالت جداسازی تصادفی و نسبت داده‌ای با درصدهای متفاوت در حالت خوشه‌بندی.



شکل ۱۱. نتایج وارون‌سازی شش نمونه اول از مجموعه داده آزمایش در دو نسبت داده‌ای (۹۸/۱/۱) و (۹۸/۱/۱)٪۵۰.

نتایجی مشابه با نسبت داده‌ای (۹۸/۱/۱) دارد نسبت (۹۸/۱/۱)٪۵۰ است. در این نسبت داده‌ای، در هیستوگرام منحنی‌های مقاومت ویژه، ٪۹۹/۷۴ نمونه‌ها مقادیر بیشتر از ۰/۷۵ و در هیستوگرام منحنی‌های فاز ۱۸/۹۴ ٪ نمونه‌ها مقادیر بین ۰/۳۶ تا ۰/۷۵ و ۸۱/۰۶ ٪ نمونه‌ها مقادیر بیشتر از ۰/۷۵ دارند. بنابراین می‌توان نتیجه گرفت که با نیمی از داده‌ها می‌توان به نتایج حاصل از کل داده‌ها رسید و این امر باعث کاهش نیاز محاسباتی و زمان آموزش می‌شود.

شکل ۱۱ نتایج وارون‌سازی شبکه برای شش نمونه اول از مجموعه داده آزمایش را برای دو حجم ۱۰۰ و ۵۰ درصد از داده‌های جداسازی شده با نسبت ۹۸/۱/۱ نشان می‌دهد. نتایج در این دو حالت بسیار نزدیک به هم و همسو با مدل واقعی یعنی مدل زمین‌گرمایی به‌کار رفته برای تولید داده است. ضخامت‌های پیش‌بینی شده برای لایه‌ها دارای انطباق خوبی با مدل واقعی هستند. بیشترین اختلاف بین مقادیر پیش‌بینی شده و مدل واقعی، در پیش‌بینی مقاومت ویژه لایه سوم یا مخزن رخ داده‌است. این لایه بین دو لایه با مقاومت ویژه پایین‌تر قرار گرفته‌است و هرچه به بیشینه مقدار مقاومت ویژه خود نزدیک می‌شود، اختلاف مشاهده‌شده بزرگ‌تر می‌شود. با احتساب محدودیت ذاتی در توان تفکیک سونداژهای MT، می‌توان قابلیت مدل‌های یادگیری عمیق در غلبه بر بخشی از توان تفکیک محدود پاسخ‌های MT را بررسی کرد که در مطالعه انجام شده توسط رحمانی جویننانی و همکاران (۲۰۲۴) تا حدودی به آن پرداخته شده‌است. ضمنا می‌توان حدس زد که اگر تعداد فرکانس‌های به‌کاررفته در تولید داده بیشتر از مقدار

فعلی (۱۳ فرکانس در بازه ۱۰۰-۰/۰۱ هرتز) باشد، یادگیری شبکه و بنابراین انطباق بین مقادیر مقاومت ویژه پیش‌بینی شده و واقعی بهبود می‌یابد.

۵ نتیجه‌گیری

در این پژوهش برای غلبه بر دو چالش نیاز به تعداد داده‌های زیاد در آموزش شبکه‌های عصبی عمیق و عدم تعمیم‌پذیری شبکه آموزش‌دیده به سبب یکسان نبودن توزیع آماری سه مجموعه داده آموزش، اعتبارسنجی و آزمایش، رویکرد کاهش مبتنی بر خوشه‌بندی داده‌ها ارائه شده‌است. عموماً جداسازی تصادفی داده‌ها برای آموزش شبکه‌های عصبی عمیق توزیع آماری یکسانی را برای سه مجموعه ایجاد نمی‌کند و رویکرد مبتنی بر خوشه‌بندی جایگزین مناسبی برای جداسازی تصادفی داده‌ها است. برای حصول اطمینان از یکسان بودن توزیع آماری سه مجموعه می‌توان از آزمون KS استفاده کرد.در خوشه‌بندی به روش K-means پس از انتخاب تعداد بهینه خوشه‌ها، از هر خوشه سهمی در سه مجموعه داده وارد می‌شود. با این تدبیر چالش یکسان نبودن توزیع آماری سه مجموعه داده برطرف می‌گردد. برای کاهش تعداد داده‌ها و بار محاسباتی ناشی از آن، از ۴۱ خوشه بهینه با درصدهای مختلف (٪۰، ٪۱۰، ٪۲۰، ٪۳۰، ٪۴۰ و ٪۵۰) از نسبت داده‌ای ۹۸/۱/۱ در قالب سه مجموعه داده آموزش، اعتبارسنجی و آزمایش، داده انتخاب می‌شود. از این داده‌ها برای آموزش یک شبکه یادگیری عمیق با معماری U-Net به منظور وارون‌سازی یک‌بعدی داده‌های مگنتوتلوریک و تعیین مشخصات یک مدل زمین‌گرمایی استفاده شده‌است. عملکرد شبکه آموزش‌دیده با معیارهای متنوعی سنجیده می‌شود. با بررسی نتایج وارون‌سازی به‌صورت کمی با کمک نمودار جعبه‌ای و کیفی با بهره‌گیری از معیار ارزیابی NSE می‌توان نتیجه گرفت که با نیمی از داده‌ها ((۹۸/۱/۱)٪۵۰) می‌توان به نتایج حاصل از اعمال کل داده‌ها (۹۸/۱/۱) رسید و این امر باعث کاهش نیاز محاسباتی و زمان آموزش می‌شود. این بررسی برای مدل ساده چندلایه و داده‌های تمیز اعمال شده است؛ با این‌وجود با اعمال روش بر داده‌های آلوده به نوفه و حاصل از مدل‌های پیچیده‌تر، مسلما تعداد داده‌های آموزشی مورد نیاز و زمان آموزش افزایش می‌یابد و قطعاً جداسازی تصادفی داده‌ها رهیافت مناسبی نخواهد بود. بنابراین پیشنهاد می‌شود صرف نظر از ساده یا پیچیده بودن مدل، بعد از جداسازی داده‌ها آزمون KS روی سه مجموعه داده اعمال و از یکسان بودن توزیع آماری آنها اطمینال حاصل شود تا تعمیم پذیری مدل با مشکل مواجه نشود. در شرایطی که توزیع‌های آماری یکسانی رخ ندهند، خوشه‌بندی می‌تواند راهکار مناسبی برای یکسان‌سازی توزیع آماری سه مجموعه و کاهش تعداد داده‌های مورد نیاز برای آموزش باشد.

منابع

خسروی، ن. ۱۳۸۵، آمار توصیفی و استنباطی. چاپ اول، انتشارات پوران پژوهش.

Alali, A., Morgan, F.D., Coles, D., 2020. Novel approach for 1D resistivity inversion using the systematically determined optimum number of layers. J Geol Geophys, 9(6), 481.

Caldwell, T.G., Bibby, H.M., Brown, C., 2004. The magnetotelluric phase tensor. Geophys. J. Int.158, 457-469.

Chen, J., Hoversten, G.M., Key, K., Nordquist, G., Cumming W., 2012. Stochastic inversion of magnetotelluric data using a sharp boundary parameterization and application to a geothermal site. Geophysics, 77(4), E265-E279.

Comeau, M.J., Becken, M., Grayver, A.V., Kaüfl, J.S., Kuvshinov, A.V., 2022. The geophysical signature of a continental intraplate volcanic system: from surface to mantle source.Earth and Planetary Science Letters 578, 117307.

Constable, S.C, Parker, R.L., Constable, C.G.,

1987. Occam's inversion: a practical algorithm for generating smooth models from electromagnetic data. Geophysics, 52, 289-300.

Fischer, G., Schnegg, P.A., Peguiron, M., LeQuang, B.V., 1981. An analytic one-dimensional magnetotelluric inversion scheme. Geophysical Journal of the Royal Astronomical Society, 67, 257-278.

Goodfellow. I., Bengio. Y., Courville. A., 2016. Deep Learning. London. The MIT Press.

Junge, A., 2011. A concept for 1D inversion of MT data using phase tensor invariants. 24. Schmucker-Weidelt-KolloquimNeustadtan der Weinstrasse, 19-23 September.

Kim, Y., Nakata, N., 2018. Geophysical inversion versus machine learning inversion in inverse problems. Leading Edge, 894-901.

Kingma, D.P., Ba, J.L., 2014. Adam: a method for stochastic optimization. In International Conference on Learning Representations.

Kodinariya, T.M., Makwana, P.R., 2013. Review on determining number of Cluster in K-Means Clustering. International Journal of Advance Research in Computer Science and Management Studies, 1 (6), 90-95.

Liao, X., Zhang, Z., Yan, Q., Shi, Z., Xu, K., Jia, D., 2022. Inversion of 1-D magnetotelluric data using CNN-LSTM hybrid network. Arabian Journal of Geosciences15, 1430.

Liao, X., Shi, Z., Zhang, Z., Yan, Q., Liu, P., 2022. 2D inversion of magnetotelluric data using deep learning technology. ActaGeophysica 70, 1047-1060,

Liu, Z., Chen, H., Ren, Z., Tang, J., Xu, Z., Chen, Y., Liu, X., 2021. Deep learning audio magnetotellurics inversion using residual-based deep convolutional neural network. Journal of Applied Geophysics,188, 104309.

Liu, W., Wang, H., Xi, Z., Zhang, R., Huang, X., 2022. Physics-driven deep learning inversion with application to magnetotelluric. Remote Sens14, 3218.

MacQueen, J. B., 1967. Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press. pp. 281–297.

Miri, H., Habibian Dehkordi, B., Payrovian, G., 2021. Oil field imaging on the Sarab Anticline, southwest of Iran, using magnetotelluric data. Journal of Petroleum Science and Engineering, 202, 108497.

Nair, V., and E. G. Hinton, 2010, Rectified linear units improve restricted Boltzmann machines: Proceedings of the 27th International Conference on International Conference on Machine Learning, 807–814.

Nash J, Sutcliffe J (1970) River flow forecasting through conceptual models part I: A discussion of principles. Journal of Hydrology, 10, 282-290.

Oh, S., Noh, K., Seol, S.J., Byun. J., 2020. Cooperative deep learning inversion of controlled-source electromagnetic data for salt delineation. Geophysics 85(4), E121-E137.

Olaniyi Muraina., I., 2022. Ideal dataset splitting ratios in machine learning algorithms: general concerns for data scientists and data analysts. 7th International mardin artuklu scientific researches conference, 496-504, www.artuklukongresi.org

Parker, R.L., Booker, J.R., 1996. Optional one-dimensional inversion and bounding of magnetotelluric apparent resistivity and phase measurements. Physics of the Earth and Planetary Interiors 98, 269-282.

Puzyrev, V., 2019. Deep learning electromagnetic inversion with convolutional neural networks. Geophys. J. Int 218, 817-832.

Rahmani Jevinani, M., Habibian Dehkordi, B., Ferguson, I.J., Rohban, M.H., 2024. Deep learning-based 1-D magnetotelluric inversion: performance comparison of architectures. Earth Science Informatics17, 1663-1677.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for

biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer, 234-241.

Segovia, M.J., Diaz, D., Selzak, K., Zuiga, F., 2021. Magnetotelluric study in the Los Lagos region (Chile) to investigate volcano-tectonic processes in the Southern Andes.Earth, Planets and Space 73(5).

Shahriari, M., Pardo, D., Picon, A., Galdran, A., Del Ser, J., Torres-Verdin, C., 2020. A deep learning approach to the inversion of borehole resistivity measurements. Computational Geosciences24, 971-994.

Smith, J.T, Booker, J.R., 1988. Magnetotelluric inversion for minimum structure. Geophysics53, 1565-1576.

Vanderkelen, I.; Van Lipzig, N.P.M.; Thiery, W. Modelling the Water Balance of Lake Victoria (East Africa)-Part 2: Future Projections. Hydrol. Earth Syst. Sci. 2018, 22, 5527–5549.

Iranian Journal of Geophysics, Vol 19, No 2, 2025, P. 6 (Research article)
DOI: 10.30499/IJG.2024.450092.1589

Reduction of the data required for training deep learning models based on clustering of the data and its application in one-dimensional magnetotelluric inversion

Mehdi Rahmani Jevinani[1] and Banafsheh Habibian Dehkordi[2*]

[1] *Ph.D. Student, Institute of Geophysics, University of Tehrn, Tehrn, Iran*
[2] *Assistant Professor, Institute of Geophysics, University of Tehran, Tehran, Iran*

(Received: 27 March 2024, Accepted: 25 September 2024)

Summary
Data-driven deep learning approaches have to deal with the challenge of generating large amounts of high-quality data, as well as the heavy computational cost and long training time imposed by it. Due to their ability to approximate complex nonlinear mapping functions, deep networks can be used effectively in geophysical inverse problems and better generalization can be achieved through deeper networks in many applications. In this research, an approach based on primary clustering of training data and assigning a certain percentage of each cluster to training, validation and test data has been used for data splitting. Kolmogorov Smirnov (KS) test has been applied to compare the distribution of three sets that are divided in this manner, and indicates that the training, validation and test data have the same distribution. A deep learning model based on modified U-Net architecture has been trained for one-dimensional inversion of magnetotelluric (MT) data, which is a highly non-linear regression problem. Supervised learning and back propagation error are used, and therefore, the inputs along with the corresponding outputs are given to the network in the form of training samples. For this purpose, a five-layer geoelectric model has been considered to simulate the conditions of a geothermal field. Using magnetotelluric forward modeling algorithm, the responses of this one-dimensional geoelectric model are analytically calculated in the frequency range of 0.01-100 Hz and in 13 frequencies that are uniformly distributed on a logarithmic scale, and total of 500000 sample data were generated. The thickness of the layers is variable and considered as part of the output. Pre-processing is done to scale the input and output variables before training and the network outputs are post-processed to be returned to the original interval. The mean square error (MSE) loss function and the Adam optimizer were used to train the network. Training is accomplished with a different amount of data separated by the mentioned method, and network performance is evaluated with some quantitative and qualitative criteria, including boxplots of Euclidean distance between true and predicted outputs and Nash Sutcliffe Efficiency coefficients. The trained network predicts the electrical resistivity and thickness of the layers from the new set of phase and apparent resistivity values. The results show that data splitting in this manner reduces the number of training data required to train the deep learning model by at least 50% without reducing the accuracy of the trained network. For noisy data and in more real scenarios, random separation is definitely not a suitable approach to form training, validation and test sets. In these conditions, the use of clustering is a suitable solution for equalizing the statistical distribution of the three sets and reducing the number of required data.

Keywords: Clustering, deep learning, inversion, magnetotelluric

[*]Corresponding author: bhabibian@ut.ac.ir