گزارش کار

رامین کهندل گرگری
کارشناسی ارشد علوم کامپیوتر
۴۰۱۳۲۴۵۵۰۲

۶ بهمن ۱۴۰۱

استاد مربوطه: دکتر برنا
درس مربوطه: یادگیری ماشین
نیم سال اول تحصیلی ۰۲-۱۴۰۱

نام مقاله A Comparative Study of Predictive Machine Learning Algorithms for COVID-19 Trends and Analysis

۱ خلاصه مقاله

در این مقاله سعی شده است با تجزیه و تحلیل بر روی داده های سازمان جهانی بهداشت (WHO) با استفاده از متد و رویکردهای مختلف رگرسیون یادگیری ماشین مانند درخت های تصمیم، LSTM و CNN یک متد پیشبینی مناسب ارائه دهد.

داده های اولیه دارای ۹۱ ورودی است که شامل داده های کشورهای مختلف با توجه به تاریخ ها به همراه موارد تایید شده، مرگ های تایید شده و موارد بهبود یافته است. (نکته با توجه به آپدیت سایت رسمی سازمان جهانی موارد بهبود یافته به دلیل تولید واکسن و استفاده گسترده از آن از دیتاست حذف شده است که با هماهنگی TA قرار بر این شد تا بر روی باقی موارد موجود متدهای لازمه را اجرا کنیم) دیتاست به نسبت ۷۰:۳۰ تقسیم شده است که در آن ۷۰ درصد برای آموزش و اعتبار سنجی و ۳۰ درصد برای آزمایش استفاده می شود. کشورهای مورد بحث در مثاله شامل کشورهای چین، کانادا و هند می باشد. شیوع بیماری کرونا در سال ۲۰۱۹ در شهر ووهان چین آغاز شد.

هدف اصلی استفاده از رویکردهای مختلف هوش مصنوعی است، که ما باید موارد تایید شده، مرگ‌های تایید شده و موارد بهبودیافته را پیش‌بینی کنیم. و علاوه بر این، از تکنیک‌های تصویرسازی مختلف برای استنباط معنادار از پیش‌بینی مدل و انجام تحلیل‌های خاص بر روی

نتایج به‌دست‌آمده استفاده شده است. مدل‌های پیش‌بینی مانند LSTM و CNN بر اساس چندین تابع خطا مانند امتیاز R2 و میانگین مربعات خطا ارزیابی می‌شوند.

۲ متدهای اجرایی

همانطور که در مقدمه اشاره شده، دیتاست استفاده شده در این مقاله به دلیل آپدیت شدن سایت سازمان جهانی بهداشت دیگر مانند قبل نیست و فقط میزان ابتلا و مرگ و فرایند تجمعی آنها را شامل است، و دیگر خبری از ستون میزان بهبودیافتگان وجود ندارد. (دلیل آن هم ساخت واکسن و همه گیری آن ذکر شده است) که با هماهنگی و صحبت با TA قرار بر این شد تا بر روی دیتاست جدید موارد را پیاده سازی کنیم که این خود باعث تفاوت خروجی خواهد شد. قبل از شروع اجرا کردن متد های مختلف ابتدا ما باید دیتاهای کشورهای کانادا، چین و هند را از دیتاست کلی جدا کرده و فقط ۳ ماه ذکر شده در کرونا که تقریبا سه ماه اول این پاندمی می باشد را برداشت کرده و فرایند رگرسیون و پیشبینی را رو آنها انجام دهیم. با بررسی دیتای موجود در دیتاست متوجه شدیم که هیچ دیتای از دست رفته یا منفی و مخرب در دیتاست وجود نداشته لذا نیازی به پاکسازی دیتاست نبود که در نوت بوک به تمامی این موارد اشاره شد. سعی شده است قبل از اجرای متدها با تصویر سازی میزان پراکندگی دیتا و چگونگی چینش دیتا در دیتاست های کشور های مختلف آشنایی پیدا کنیم، که این کار را با استفاده از لایبری های معروف پایتون جهت ساخت پلات استفاده کردیم. (matplotlib و seaborn) تقسیم بندی دیتا جهت تست و آموزش به نسبت ۷۰ به ۳۰ در تمامی مراحل انجام شده است.

درخت درخت تصمیم در دسته الگوریتم های یادگیری ماشینی نظارت شده قرار می گیرد و می توان از آن برای متغیرهای خروجی پیوسته و دسته بندی استفاده کرد.در این روش از یک flowchart تصمیم گیری مانند ساختار درختی برای تصمیم گیری استفاده می شود. شاخه ها معمولاً نتیجه گره و را نشان می دهند گره ها دارای یکی از این موارد هستند:

شرایط (گره های تصمیم گیری)

نتایج (گره های پایانی)

رگرسیون درخت تصمیم ویژگی‌های یک شی را مشاهده می‌کند و مدلی را در ساختار درخت آموزش می‌دهد تا داده‌ها را در آینده پیش‌بینی کند تا خروجی پیوسته معنی‌داری تولید کند. خروجی ای که گسسته نیست به عنوان خروجی پیوسته شناخته می شود، یعنی فقط مجموعه ای گسسته از اعداد یا مقادیر نشان داده نمی شود. جهت پیاده سازی درخت تصمیم در پایتون از لایبری های sklearn و DecisionTreeRegressor استفاده کرده ایم که تمامی مراحل اجرای آن داخل نوت بوک آورده شده است.

CNN CNN ها به طور کللی برای داده های دو بعدی طراحی می شدند، اما می توان از آنها برای مدل سازی پیش بینی سری های زمانی تک متغیره نیز استفاده کرد. مجموعه‌های زمانی تک متغیره مجموعه‌ای از گروه های منفرد مشاهدات هستند که در آن یک ترتیب زمانی وجود دارد که مدل ایجاد شده برای یادگیری و درک از سری مشاهدات گذشته برای پیش‌بینی مقدار بعدی در دنباله مورد نیاز است.

پس آماده سازی یک سری تک متغیره قبل از مدل سازی آن ضروری است. برای تهیه آن باید دنباله داده های خود را به الگوهای ورودی/خروجی متعددی به نام نمونه تقسیم کنیم در مورد ما، از سه مرحله زمانی به عنوان ورودی و یک گام زمانی به عنوان خروجی برای پیش‌بینی یک مرحله‌ای که در حال یادگیری است استفاده می‌کنیم. گام بعدی توسعه یک مدل CNN است، یک CNN یک بعدی یک مدل CNN است که دارای یک لایه پنهان کانولوشنی است که روی یک دنباله ۱ بعدی عمل می کند. اگر دنباله‌های ورودی طولانی داشته باشیم، یک لایه کانولوشنال دیگر دنبال می‌شود و سپس یک لایه ادغام که ویژگی اصلی آن استخراج مهم‌ترین ویژگی از داده‌ها است، دنبال می‌شود. سپس با لایه کاملاً متصل متراکم دنبال می شود که برای تفسیر ویژگی های استخراج شده از بخش کانولوشنال مدل استفاده می شود.سپس یک لایه مسطح بین لایه های کانولوشن و لایه متراکم استفاده می شود تا نقشه های ویژگی را به یک بردار تک بعدی کاهش دهد. می‌توانیم لایه‌ها را اضافه یا کم کنیم تا ببینیم چه نوع تأثیری بر داده‌های آموزشی با استفاده از برخی معیارهای عملکرد مانند مربع‌های R دارد. سپس داده های تست از طریق این مدل برای پیش بینی عدد بعدی در دنباله منتقل می شوند.

برای پیاده سازی CNN در پایتون من از لایبری و داکیومنتیشن tensorflow استفاده کرده ام که تمامی موارد لازم برای ایجاد این متد را در اختیارمان قرار می دهد و تمامی مراحل در داخل نوت بوک آورده شده است.

نکته : موارد ذکر تمامی آن چیزی بود که مقاله جهت ایجاد یک شبکه عصبی در اختیار ما قرار داده که همه ما میدانیم که موارد بیشتر از این می باشد و این پارامترها لازمه اصلی مسئله جهت پیاده سازی یک شبکه عصبی مشابه خود مقاله می باشند. مواردی مانند تعداد epoch و غیره که در ساخت شبکه و کارکرد آن اهمیت بسیاری دارند مواردی هستند که هیچ اشاره ای در مقاله نشده و من صرفا از یک سری آزمون و خطا و مقادیر پیش فرض جهت ایجاد آن استفاده کرده ام که قطعا با متد اصلی متفاوت خواهد بود. در صورت نیاز میتوانم تمامی این پارامترهایی که لازمه اصلی بوده و هیپ اشاره ای به آنها نشده است را نیز ارائه دهم.

تمامی موارد گفته شده برای روش LSTM نیز صادق است.

LSTM برای مدیریت وابستگی‌های توالی، نوع قدرتمندی از شبکه عصبی طراحی شده است که به آن شبکه عصبی بازگشتی (RNN) می‌گویند. LSTM نوعی RNN است که می تواند برای حل مسائل پیش بینی سری های زمانی استفاده شود که بر مشکل محو شوندگی گرادیان غلبه می‌کند و با استفاده از Back Propagation در طول زمان آموزش داده می‌شود. همچنین برای رسیدگی به مشکلات توالی پیچیده در یادگیری ماشین استفاده می شود.

LSTM .یک بلوک از گیت هایی تشکیل شده است که وضعیت و خروجی بلوک را مدیریت می کنند هر گیت درون یک بلوک بر اساس توالی ورودی کار می‌کند از یک واحد فعال‌سازی سیگموئید برای کنترل اینکه آیا تغییر حالت و اضافه کردن اطلاعات در بلوک را مشروط می‌کند یا خیر، استفاده می‌کند. LSTM از سه نوع گیت زیر در یک واحد تشکیل شده است:

Forget Gate: برای تصمیم گیری اینکه چه اطلاعاتی را نگه دارید و چه چیزی را از بلوک دور بریزید استفاده می شود.

گیت ورودی: از یک شرط برای تصمیم گیری برای به روز رسانی حافظه از ورودی استفاده می کند.

گیت خروجی: از یک شرط برای تصمیم گیری برای خروجی بر اساس ورودی و حافظه بلوک استفاده می کند.

هر واحد مانند ماشینی است که در آن گیت ها وزن هایی دارند که در طول تمرین یاد می‌گیرند.
برای پیاده سازی LSTM در پایتون من از لایبری و داکیومنتیشن tensorflowاستفاده کرده ام که تمامی موارد لازم برای ایجاد این متد را در اختیارمان قرار می دهد و تمامی مراحل در داخل نوت بوک آورده شده است که نبود اطلاعات کافی خود را در این روش واضحا نشان می دهد.

۳ متدهای دیگر

پس از اجرایی تمامی متدهای ذکر شده در مقاله از آنجایی اینکه نتوانستیم به علت نبود اطلاعات کافی در اجرای توابع لازم برای روش های مختلف که قبلا هم ذکر شد، سعی کردم روش رگرسیون خطی و چند جمله ای را نیز بر روی این دیتاست ها اجرا کنم تا حداقل یک پیشبینی مناسب بدست بیاوریم.
همانطور که در کدهای اجرایی هم پیداست پس از اجرای رگرسیون خطی که جواب آنچنان مناسبی برای موارد ما ارائه نمیداد، رگرسیون چند جمله ای را با آزمون و خطا درجه چند جمله ای آن بین اعداد ۲ تا ۹ اجرا کردیم تا بتوان بهینه ترین جواب از این نظر را ارائه داد.
جهت پیاده سازی این رگرسیون ها در پایتون از لایبری های sklearn و PolynomialFeatures و LinearRegression استفاده کرده ایم که تمامی مراحل اجرای آن داخل نوت بوک آورده شده است.

۴ ارزیابی نتایج

دراین مقاله برای ارزیابی از نتایج از روش R2 Score استفاده می کند که با استفاده از کتابخانه sklearn به راحتی قابل دسترسی است.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

از آنجایی که اجرای کاملا مشابه روش ها مانند مقاله به دلیل اطلاعات کم امکان پذیر نبود، نتایج اجرایی این آنالیز برای تمامی دیتاست ها و تمامی متدهای اجرا شده صرفا داخل نوت بوک ذکر شده و از تکرار آنها در گزارش کار خودداری شده است.

۵ جمع بندی

در این تحقیق تجربی، سعی شده بود با دانش و بینش، الگوهایی را از داده‌های real time جهانی برای پاندمی COVID-19 ارائه شود.

پس از ادغام موفقیت‌آمیز داده‌ها با الگوریتم‌ها، تجزیه و تحلیل‌ها با استفاده از کتابخانه‌های پایتون مانند matplotlib، seaborn، plotly، bokeh و altair تجسم شدند. نمایش های نموداری مانند نمودارهای خطی و نمودارهای میله ای روند شناسایی شده در داده های سطح کشور/ایالت را توضیح می دهند. برای موضوع همه‌گیر کووید-۱۹، می‌توان نتیجه گرفت که این مدل‌های پیش‌بینی می‌توانند در ارائه یک ایده ملموس و آمار مفید برای پشتیبانی از مدل بلادرنگ ما مفید باشند. علاوه بر این، می توان از همین مدل برای تعیین پیش از ورود و اقداماتی که برای کاهش تأثیر هر بیماری همه گیر یا همه گیری قریب الوقوع انجام می شود، استفاده کرد. پس از انجام آموزش و تجزیه و تحلیل دقیق بر روی داده‌ها، نتایج پیش‌بینی‌شده و real time بسیار مشابه بودند و روند متقابلی بین این دو مشاهده شد که منجر به استنباط قطعی فوق‌الذکر شد. این مدل دارای هزار ظرفیت در زمینه هوش مصنوعی یادگیری ماشین است و پتانسیل پیشرفت هایی را در خود جای داده است که در آینده می توان آنها را نیز بررسی کرد. دامنه این تحقیق را می توان با افزودن تعداد کشورها و سطح بندی اطلاعات به منظور افزایش کارایی این مطالعه تطبیقی گسترش داد. الگوریتم های مورد استفاده برای این مطالعه را می توان بر روی بیش از یک معیار عملکرد ارزیابی کرد که می تواند میانگین مربعات خطا (MSE) ،واریانس و غیره باشد دیگر رگرسیون های چشمگیر مانند Random Forest، XGBoost و ARIMA را می توان برای گسترش مطالعه مقایسه ای اضافه کرد.

در انتها سعی شد با اجرای یک روش خارج از منبع اصلی، به نام رگرسیون خطی و چند جمله ای یک پیشبینی نسبتا مناسب جهت ارائه و نمایش به عنوان خروجی این پروسه و فعالیت بدست آورد.

امیدوارم با وجود تمامی مواردی که داخل مقاله ذکر نشده بود و اجرای مشابه با آن را تقریبا غیر ممکن کرده بود توانسته باشم یک فایل مناسب جهت ارائه آماده کرده باشم.

مراجع

[۱] A. Kunjir, D. Joshi, R. Chadha, T. Wadiwala and V. Trikha, ”A Comparative Study of Predictive Machine Learning Algorithms for COVID-19 Trends and Analysis,” 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 2020, pp. 3407-3412, doi: 10.1109/SMC42975.2020.9282953.