نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران
دانشکده مهندسی کامپیوتر، دانشگاه شهید بهشتی، تهران، ۱۵-۱۳ اسفندماه ۱۳۹۲

پیش‌بینی انتشار بیماری‌های واگیردار از طریق تحلیل شبکه اجتماعی توییتر

سهیلامولائی[۱]، محمدخوانساری[۲]، مصطفی صالحی[۳]، هادی ویسی[۴]

[۱] دانشکده علوم و فنون نوین، دانشگاه تهران، تهران،
soheila.molaei@ut.ac.ir

[۲] استادیار، گروه علوم و فناوری شبکه، دانشکده علوم و فنوین نوین، تهران
m.khansari@ut.ac.ir

[۳] استادیار، گروه علوم و فناوری شبکه، دانشکده علوم و فنوین نوین، تهران
mostafa_salehi@ut.ac.ir

[۴] استادیار، گروه علوم و فناوری شبکه، دانشکده علوم و فنوین نوین، تهران
h.veisi@ut.ac.ir

چکیده

بیماری فصلی آنفلوآنزا هر ساله باعث مرگ ۵۰۰.۰۰۰ نفر در جهان می‌شود. بنابراین جلوگیری از این بیماری و بیماری‌های واگیردار مشابه از اهمیت بالایی برخورداراست. همان‌طور که مطالعات نشان می‌دهد، در صورت تشخیص زودهنگام، می‌توان از بسیاری از بیماری‌های واگیردار جلوگیری کرد. ازاین‌رو، پیش‌بینی شیوع بیماری‌های واگیر نقش مهمی در کاهش خسارات ناشی از آنها دارد. مرکز کنترل و پیشگیری بیماری (CDC) به صورت سنتی داده بیماری شبه آنفلوآنزا (ILI) را جمع‌آوری می‌کند که معمولاً فاصله زمانی تشخیص بیماری تا تهیه گزارش ILI، یک تا دو هفته است. کاهش این زمان، منجر به پیش‌بینی زودتر شیوع بیماری و کاهش هزینه‌های آن می‌باشد. در این مقاله با استفاده از روش‌های یادگیری ماشین، روش‌های جدیدی برای پیش‌بینی داده بیماری شبه آنفلوآنزا مبتنی بر تحلیل داده‌های شبکه اجتماعی توییترارائه شده است. روش‌های پیشنهادی از مدل رگرسیون خطی با ورودی خارجی و مدل سری زمانی با شبکه عصبی برای پیش‌بینی داده بیماری شبه آنفلوآنزا استفاده می‌کنند. ارزیابی‌های انجام شده بر روی داده‌های سال ۲۰۰۹ تا ۲۰۱۰ شبکه توییتر نشان می‌دهند که از طریق روش‌های پیشنهادی امکان پیش‌بینی شیوع بیماری، دو تا چهار هفته زودتر از CDC، فراهم می‌شود. نتایج حاصل نشان می‌دهد که می‌توان توسط شبکه عصبی بیماری آنفولانزا را با خطای کمتر از ۵٪ پیش‌بینی کرد.

کلمات کلیدی

پیش‌بینی بیماری واگیردار، آنفلوآنزا، توییتر، سری زمانی با شبکه عصبی، رگرسیون .

۱- مقدمه

بیماری فصلی آنفلوآنزا سالیانه باعث مرگ ۵۰۰.۰۰۰ نفر در جهان می‌شود [۱]. بین سال‌های ۱۹۱۸ تا ۱۹۲۰ آنفلوآنزای اسپانیایی باعث مرگ ۲۰ تا ۱۰۰ میلیون نفر در دنیا شد[۲]. بنابراین کم کردن تأثیر این بیماری‌ها مثل آنفلوآنزای گونهی H1N1 بسیار مهم شد. یکی از راه‌ها ردگیری و پیش‌بینی پخش بیماری در یک جمعیت است. تحقیقات نشان می‌دهد که بسیاری از این بیماری‌ها درصورت تشخیص زودهنگام قابل کنترل هستند [۳].

مرکز کنترل و پیشگیری از بیماری‌ها (CDC)[۱] [۴]، که بیماری‌های شبه آنفلوآنزا (ILI)[۲] را با جمع‌آوری داده نظارت می‌کند، به صورت هفتگی، ماهانه و سالانه گزارش خود را تهیه می‌کند. این

مرکز در زمینه پزشکی بسیار معتبر است ولی تا بیماری توسط پزشک تشخیص داده شود و گزارش‌ها منتشر شود، زمانی بین یک یا دو هفته طول می‌کشد. این در حالی است دولت‌ها باید در زودترین زمان مطلع شوند تا بتوانند به صورت کارا از خطر همه‌گیری بیماری جلوگیری کنند. لازمه این مهم، داشتن روش‌های سریع و کارا برای پیش‌بینی بروز بیماری‌های واگیردار همچون آنفلوآنزا است.

سیستم‌های نظارتی مختلفی برای پایش رفتارهای مرتبط با سلامتی و استفاده از آن‌ها برای تشخیص فعالیت ویروس آنفلوآنزا پیشنهاد شده است. برای مثال، می‌توان به بررسی تعداد داروهای فروخته شده، پرونده بیماران که تحت نظر پزشک برای آنفلوآنزا قرار گرفته‌اند، اشاره کرد. گوگل برای تحلیل روند شیوع بیماری آنفلوآنزا ابزاری را معرفی کرده است که بر مبنای رشد میزان ثبت وقایع جستجوهای مرتبط با بیماری آنفلوآنزا کار می‌کند [۵].

رشد شبکه‌های اجتماعی برخط در سال‌های اخیر، آنها را به عنوان پرکاربردترین کاربردهای اینترنت تبدیل کرده است. برای مثال توییتر یک شبکه‌ی اجتماعی است که به کاربران اجازه می‌دهد تا ۱۴۰ حرف، پیام متنی را که توییت نامیده می‌شود، ارسال کنند[۶].

ایده کلی این است که اینگونه سایت‌ها می توانند برای نظارت بر روند سلامتی از جمله شیوع بیماری‌های واگیردار همچون آنفلوآنزا مورد استفاده قرار گیرند. بطور کلی در شیوع بیماری روش‌های یادگیری‌ماشین و داده‌کاوی و مدل‌های احتمالی مانند بردار پشتیبان و رگرسیون[۷,۸] و فیلد تصادفی شرطی[۳] [۹,۱۰,۱۱] استفاده شده است. روند کلی کار اینگونه است که در ابتدا احتیاج به مجموعه داده واقعی از توییتر است. داده‌ها را می‌توان با نوشتن برنامه‌ای تحت وب (و بکارگیریAPIهای توییتر) از روی این سایت جمع‌آوری کرد[۱۲]. در ادامه باید به توییت‌ها برچسب «بیمار»یا «سالم» زده شود. یک سری از توییت‌ها داده‌های آموزشی در نظر گرفته می‌شوند و در این داده‌های آموزشی یادگیری حاصل می‌شود. یادگیری حاصله در واقع به کلمات وزن می‌دهد و از روی وزن‌ها تشخیص می‌دهد که برچسب بیمار یا سالم مناسب‌تر است. بعد از یادگیری می‌توان بقیه داده‌های باقی‌مانده یا داده‌های تست را با استفاده از یادگیری حاصل شده برچسب بیمار یا سالم بزنیم. در بعضی مقالات علاوه بر این مکان کاربر را نیز در نظر گرفته‌اند، چون ارتباطات انسانی فاکتور مهمی در بیماری‌های عفونی است [۹,۱۰,۱۱].

برچسب بیمار یا سالم زدن به هر توییت نیاز به تعداد قابل توجهی داده آموزشی برچسب خورده دارد تا بتوان مابقی توییت‌ها را با روش‌های یادگیری ماشین برچسب زد که مشخصاً فرایند زمان‌گیری است [۱۱]. همچنین روش‌های دیگرکه از برچسب استفاده نمی‌کنند نیز خطای نسبتاً بالاتری دارند [۱۳]. روش‌های موجود [۱۳] با استفاده از توییت‌های هر هفته به پیش‌بینی بیماری در همان هفته می‌پردازند که در واقع یک تا دو هفته از گزارش CDC جلوتر است. در این روش‌ها فقط به پیش‌بینی CDC پرداخته می‌شود و نهایتاً مقایسه با CDC انجام می‌گیرد.

در این مقاله سعی شده است که روش‌های دقیق‌تر (با درصد خطای پایین‌تر) و در عین حال در زمان کمتر و بدون نیاز به برچسب‌گذاری ارائه شود. همچنین اگر بتوان هفته‌های جلوتری از CDC را پیش بینی کرد می‌توان خطر شیوع آنفلوآنزا را در زمان زودتری پیش‌گیری کرد که روش‌های پیشنهادی دو تا چهار هفته زودتر از CDC پیش‌بینی می‌کنند که در واقع یک تا دو هفته جلوتر از نزدیکترین روش به ما در این حوزه است [۱۳]. همچنین اگر بتوان مرجع مقایسه را افزایش داد درستی روش‌های موجود ملموس‌تر خواهد بود. با توجه به این مطلب علاوه بر CDC، در این مقاله ابزار گوگل برای پیش‌بینی آنفلوآنزا نیز به عنوان مرجع مقایسه در نظر گرفته شده است. نتایج نشان می‌دهند که روش‌های پیشنهادی برای هر دو مرجع CDC و گوگل خطای کمتری نسبت به روش‌های موجود دارند.

۲- کارهای مرتبط

بسیاری از مطالعات بر روی شبکه‌های اجتماعی‌ای مانند فیسبوک، توییتر، فلیکر، لینکداین، ویکی‌پدیا، یوتیوب و غیره صورت گرفته است. در واقع امروزه رسانه‌های اجتماعی فرصتی برای بهره‌برداری از داده‌ها و پیش‌بینی نتایج دنیای واقعی با ساخت مدل‌ها و بدست آوردن رفتار انسان‌ها شده‌اند. درادامه به مروربرخی از تحقیقات پرداخته شده است.

توییتر برای اعلان‌های بی‌درنگ مانند شرایط اضطراری آتش‌سوزی‌هایی با مقیاس بزرگ و ترافیک زنده می‌تواند استفاده شود [۱۴]. این طبیعت توزیعی توییتر باعث می‌شود ابزار قدرتمندی برای روزنامه‌نگاری شهری شود. همچنین از توییت‌های توییتر برای استفاده در تشخیص خلق و خوی ملی نیز استفاده می‌شود چون هر توییت خلق و خوی نویسنده را تشریح می‌کند [۱۵]. گینزبرگ و دیگران [۱۶] رویکردی برای پیش‌بینی روند آنفلوآنزا با استفاده از جستجوهای انجام شده در گوگل ارائه داده است که همبستگی خطی بین تعداد جستجوها و تعداد بیمارانی که توسط پزشکان ویزیت شده‌اند، استخراج کرده است.

لامپوز وکریستیانینی [۱۷] بر روی شیوع آنفلوآنزای گونه‌ی H1N1 در انگلیس با بررسی توییت‌های ارسالی در ۲۴ هفته تحقیق کردند. در این پژوهش جملاتی که علائم بیماری را هشدار می‌دادند بررسی شده و در نهایت به یک امتیازآنفلوآنزا تبدیل شده است. با مقایسه این امتیاز با آژانس بهداشت و درمان بیش از ۹۵ درصد همبستگی خطی بدست آمده است که قابل توجه است.گینزبرگ و دیگران[۱۶] پیام‌های گمراه‌کننده در توییتر را مورد بررسی قرار داده و نشان دادند که تنها بخش کمی از کلمات کلیدی مرتبط با آنفلوآنزا

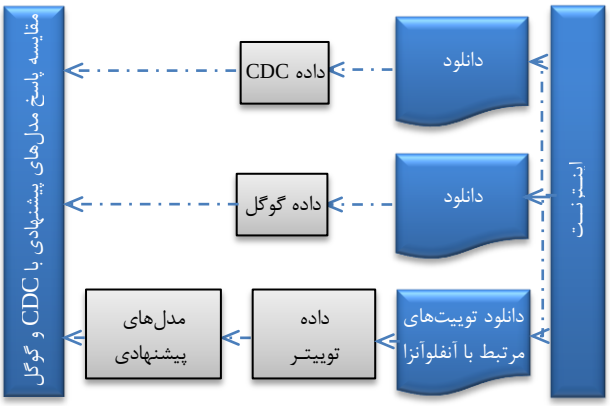
می‌توانند در پیش‌بینی ارتباط به کار آیند. پال و دردز[۱۸] بیش از نیم میلیون توییت‌های مرتبط با سلامت را بررسی کردند و کلماتی مرتبط با بسیاری از بیماری‌ها مثل آلرژی و چاقی و بی‌خوابی را کشف کردند که این‌ها برای پیگیری بیماری در طول زمان می‌توانند مورد استفاده قرار گیرد. ولوسو و فراز[۱۹] نشان دادند که بیماری عفونی دانگ که از طریق نوعی پشه در برزیل منتقل می‌شود و علت عمده مرگ‌ومیر درمناطق گرمسیری و نیمه‌گرمسیری از جمله برزیل است، چگونه می‌تواند از طریق توییتر پیگیری شود.

آچرکار و همکارانش[۱۳] ابزاری بی‌درنگ به نام SNEFT[۴][۲۰] برای کنترل روند آنفلوآنزا از طریق شبکه اجتماعی ارائه کردندکه درواقع هدف آن ساخت ابزاری برای کنترل بیماری آنفلوآنزا از طریق اطلاعات سایت‌های فیسبوک و توییتر است. برای این منظور از لیستی از کلمات کلیدی‌ای مانند آنفلوآنزا و H1N1 و پیام‌های مرتبط با بیماری و همچنین اطلاعات کاربر (مثل زمان و مکان) برای پیاده‌سازی یک چارچوب برای پیش‌بینی بیماری با استفاده از داده‌های جمع‌آوری شده از توییت‌ها استفاده شده است. آن‌ها توییت‌های سال ۲۰۰۹-۲۰۱۰ را که زمان اوج آنفلوآنزای خوکی بود را تحلیل کرده و مطالعه‌ای روی آنفلوآنزای فصلی با استفاده از داده‌کاوی و مدل رگرسیون انجام داده‌اند. همچنین آچرکار و دیگران[۲۱] پیشنهاد می‌دهند که چون گرایش‌های برخط اجتماعی با تغییر زمان، مکان جغرافیایی و جمعیتی مانند سن و جنسیت می‌تواند تغییرکند، بهترین سن برای پیش‌بینی آنفلوآنزا بین گروه‌های سنی ۵-۲۴ و ۲۵-۴۹ است.

مشکل روش‌های موجود این است که به پیش‌بینی همین هفته می‌پردازند. هرچه پیش‌بینی سریع‌تر صورت گیرد می‌توان از وقوع بیماری‌ها با سرعت بیشتری پیشگیری کرد. برای این مشکل روش‌های پیشنهادی با استفاده هفته قبل یا دو هفته قبل به پیش‌بینی می‌پردازند در نتیجه یک تا دو هفته جلوتر پیش‌بینی می‌شود. مشکل دیگر این است وقتی مرجع مقایسه فقط CDC باشد نمی‌توان از درستی روش‌های ارائه شده خیلی مطمئن بود. چون گوگل نیز به پیش‌بینی بیماری آنفلونزا می‌پردازد برای حل این مشکل داده گوگل نیز به‌کار گرفته شد یعنی از توییت‌ها برای پیش‌بینی گوگل استفاده شد. در این صورت روش‌های ارائه شده قابلیت اطمینان بیشتری پیدا خواهدکرد. بعد از پیاده‌سازی و مقایسه مشخص شد که درصد خطای مطلق برای CDC و گوگل به طور میانگین هشت تا هجده درصد کاهش یافته است.

مراحل یک سیستم پیش‌بینی بیماری

به دلیل اینکه توییت‌ها غیرعددی هستند، برای پیش‌بینی بیماری آنفلوآنزا در ابتدا احتیاج به استخراج داده از توییت‌ها پس از جمع‌آوری آن‌هاست. علاوه بر این، برای اثبات درستی نتایج احتیاج به داده مرجع است تا پیش‌بینی و مقایسه صورت گیرد و در نهایت خطا محاسبه گردد. در این قسمت انواع داده مورد استفاده، روش جمع‌آوری و تحلیل شرح داده شده است. شکل (۱) روند کلی اقدامات انجام گرفته برای پیش‌بینی بیماری واگیردار آنفلوآنزا را نشان می‌دهد.



شکل (۱): مراحل پیش‌بینی بیماری واگیردار آنفلوآنزا

۲-۱- داده‌های مورد استفاده

در حالت کلی سه قسم داده مورد استفاده قرار می گیردکه داده اول به عنوان ورودی برای پیش‌بینی و داده دوم و سوم به عنوان داده مرجع و خروجی است:

۱- داده‌های جمع‌آوری شده از توییتر یا داده شبکه اجتماعی برخط
۲- داده‌های مرکز کنترل و پیشگیری از بیماری‌ها
۳- داده‌های روند شیوع آنفلوآنزای گوگل

در چارچوب SNEFT راه‌اندازی شده لیستی از کلمات مهم مرتبط با آنفلوآنزا از جمله ’influenza‘ ، ’H1N1‘، ’Flu‘، ’Swine Flu‘ درون توییت‌های کاربران جستجو شده و توییت‌های مرتبط با آن جمع‌آوری شده است[۱۳].

داده‌های مربوط به بیماران مبتلا به آنفلوآنزا از سایت مرکز کنترل و پیشگیری از بیماری‌ها[۴]، داده‌های روند شیوع بیماری آنفلوآنزای گوگل نیز از سایت گوگل[۲۲] دریافت و ذخیره گردید.

۲-۲- آنالیز داده توییتر

از تاریخ هجدهم اکتبر ۲۰۰۹ تا تاریخ سی و یکم اکتبر ۲۰۱۰ توییت‌های کاربران که کلمات مرتبط با آنفلوآنزا در آن‌ها بود مورد بررسی قرار گرفت و توییت‌های مرتبط با امریکا جمع‌آوری شد. بنابراین ۴.۷ میلیون توییت از ۱.۵ میلیون کاربرتوییتر بررسی شده است[۱۳].

در مرحله بعدی تکرارتوییت[۵]ها و توییت‌های بعدی از کاربرمشابه از بین توییت‌ها حذف شده است. ازبین توییت‌های باقی‌مانده، توییت‌هایی که کلماتی چون واکسینه کردن و نظیر آن را داشتند نیز از

بین توییت‌ها حذف شدند. همان‌گونه که در جدول (۱) نشان داده شده است، با پردازش‌های بیان شده، همبستگی بین داده CDC و توییت‌های جمع‌آوری شده بیشتر می‌شود [۱۲].

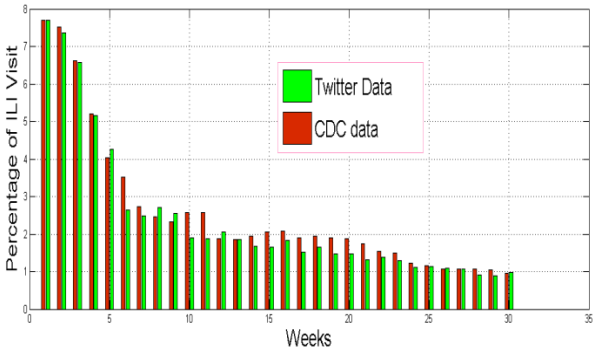
مدت زمان سپری سندروم(SET)[۶] : مدت زمانی که لازم است توییت‌های کاربر بعد از احتساب اولین توییتش در نظر گرفته نشود.

جدول(۱) : همبستگی بین داده توییتر و CDC[۱۲]

	5-fold	10-fold
مجموعه داده	ضریب همبستگی	خطای نسبی مطلق	ضریب همبستگی
توییت بدون تکرارتوییت وبدون ذکر واکسینه کردن و SET یک هفته	۰.۹۸۸۶	%۱۶.۶۲	۰.۹۸۸۵

بصورت کلی سی هفته داده پس از جمع آوری وجود دارد. با مقایسه داده CDC با توییتر مشخص شد که بهترین همبستگی مربوط به مجموعه داده بدون تکرارتوییت و SET یک هفته و بدون ذکر کلماتی مثل واکسینه کردن است.

در جدول(۱) ضریب همبستگی و خطای نسبی مطلق توسط رگرسیون خطی بین داده توییتر و CDC نشان داده شده است. ضریب همبستگی یک معیار آماری است که ارتباط بین دو متغیر را نشان می‌دهد و بین ۱- تا ۱ است. علامت ضریب همبستگی نشان دهنده این است که این دو به صورت مثبت یا منفی با هم ارتباط دارند و مقدار آن نشان دهنده این است که چقدر با هم مرتبطند. خطای نسبی مطلق هر چه کمتر باشد داده‌ها باهم متناسب‌ترند و در واقع معیاری است که نشان می‌دهد مدلی پیش‌بینی شده تا چه اندازه صحیح است.برای بدست آوردن ضریب همبستگی رگرسیون خطی با اعتبارسنجی[۷]و تقسیم داده‌ها (n-fold) به ۵ و ۱۰ قسمت مورد مقایسه قرار گرفته است.



شکل(۲) : درصد بیماران ملاقات شده و مجموعه داده توییتر

شکل (۲) نمودار ارتباط بین بیماران ملاقات شده توسط CDC و مجموعه داده توییتر بعد از حذف تکرارتوییت و با در نظر گرفته SET یک هفته می‌باشد. داده توییتر در هر هفته تقسیم بر ۵۸۶۰ شده است که با داده CDC در یک محدوده باشند. همانطور که شکل نشان می‌دهد این دو مجموعه داده همبستگی بالایی دارند.

۳- پیش‌بینی بیماری آنفلوآنزا

در این بخش روش‌های پیشنهادی شرح داده می‌شود. این روش‌ها پیاده‌سازی شده و نتایج روش موجود[۱۳] و روش‌های پیشنهادی مورد مقایسه قرار می‌گیرد.

۳-۱- مدل رگرسیون با ورودی خارجی[۸]

مدل رگرسیون خطی با ورودی خارجی ولی با یک ورودی و خروجی است . در این مدل می‌توان خروجی جدیدی مثل $Y_i(t)$ را ایجاد کرد و نهایتاً این خروجی به عنوان ورودی برای پیش‌بینی هفته بعدی استفاده شود که در فرمول (۱) آمده است.

$Y_1(t)$ نشان‌دهنده درصد افراد بیمار در هفته t می‌باشد که توسط CDC اعلام می‌شود.

$U(t)$ نشان‌دهنده توییت‌های مرتبط با آنفلوآنزا بدون تکرارتوییت وبدون ذکر واکسینه کردن و SET یک هفته است.

$Y_i(t)$ نشان دهنده ضرب توییت‌ها و افراد بیمار در هفته t است که این Y توسط توییت های دو هفته قبل پیش‌بینی می‌شود و سپس در این فرمول استفاده می‌گردد.

$e(t)$ متغیرهای تصادفی غیروابسته است.

$A_i(Z)$ و $B(Z)$ ضرایب عددی رگرسیون است.

در این مقاله، n=2 وo=3 در نظر گرفته شده است.

$$Y_1(t) = \sum_{j=1}^{n} A_i(Z)Y_i(t-j) + \sum_{k=2}^{o} B(Z)U(t-k) + e(t) \quad (۱)$$

همچنین می‌توان ضرب توییت‌ها و افراد بیمار را به عنوان ورودی و ویژگی اضافه کرد یعنی دو ورودی و یک خروجی که در فرمول (۲) نشان داده شده است.

$$Y(t) = \sum_{i=2}^{n} B(z)U(t-j) + \sum_{j=2}^{o} C(z)U_1(t-j) + e(t) \quad (۲)$$

$Y(t)$ نشان‌دهنده درصد افراد بیمار در هفته t می‌باشد که توسط CDC اعلام می‌شود.

$U(t)$ نشان‌دهنده توییت‌های مرتبط با آنفلوآنزا بدون تکرارتوییت وبدون ذکر واکسینه کردن و SET یک هفته است

$U_1(t)$ نشان دهنده ضرب توییت‌ها و افراد بیمار است.

$e(t)$ متغیرهای تصادفی غیروابسته است.

$B(Z)$ و $C(Z)$ ضرایب عددی رگرسیون است.

در اینجا n=3 وo=3 در نظر گرفته شده است.

برای رسیدن به پاسخ دقیق‌تر از اعتبارسنجی استفاده می‌کنیم. در این مقاله، اعتبارسنجی 5-fold را در نظر گرفتیم، بدین معنا که

داده به پنج قسمت با اندازه تقریباً مساوی تقسیم شده، چهار قسمت به عنوان داده آموزش و بقیه به عنوان داده تست در نظر گرفته می‌شود و خطای جذر میانگین مربع‌ها (RMSE)[۹] و خطای نسبی مطلق (RAE)[۱۰] محاسبه می‌شود. ضرایب رگرسیون از روی ده داده اول محاسبه شده است و از روی این ضرایب پیش‌بینی برای داده‌های بعدی صورت می‌گیرد. نتایج حاصل در جدول (۲) نشان داده شده است. با استفاده از رگرسیون با ورودی خارجی و اضافه کردن ویژگی جدید به عنوان ورودی یا خروجی مشخص شد که می‌توان دو هفته جلوتر از روش‌های قبلی را با بهبود چهار درصدی خطای مطلق(RAE) پیش‌بینی کرد.

۳-۲- مدل سری زمانی با شبکه عصبی

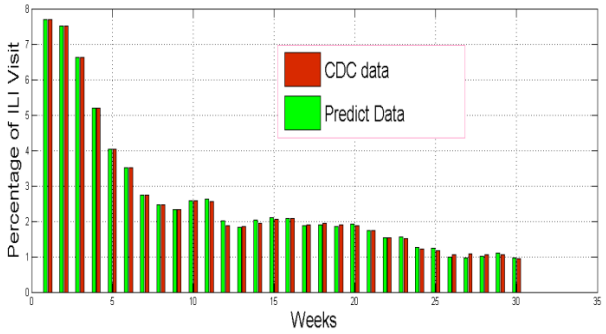
علاوه بر مدل‌های پیشین، در این مقاله سعی شده است که مدل پیچیده‌تری برای پایین آوردن خطا ارائه شود. بدین منظور مدل سری زمانی یا به عبارتی مدل رگرسیون غیرخطی با ورودی خارجی[۱۱] با شبکه عصبی پیاده‌سازی شده است. برای این کار از شبکه عصبی پرسپترون چندلایه (MLP) با سه لایه استفاده شده است. در لایه اول، دو نرون که هرکدام بیانگر ۱۰ هفته قبل ورودی (توییت) داده و بازخورد از خروجی (CDC یا گوگل) پیش‌بینی شده به ورودی هستند، استفاده شده و در لایه پنهان ۳۰ نرون در نظرگرفته شده است. خروجی شبکه شامل یک نرون است که بیانگر پیش‌بینی داده CDC یا گوگل در هفته جاری است. در آموزش شبکه از روش اعتبارسنجی استفاده شده است و به صورت مشابه با نتایج مدل قبلی، 5-fold به کار گرفته شده است. نتایج این روش نیز درجدول (۲) آمده است.

همان‌گونه که نتایج نشان می‌دهد این روش بهترین پاسخ را بین روش‌ها دارد و خطای RMSE و RAE را روی داده CDC تقریباً هشت درصد و روی داده گوگل هجده درصد می‌دهد.

جدول(۲) : نتایج پیاده‌سازی مدل‌ها روی داده CDC

فرمول	RMSE	RMSE(درصد)	RAE(درصد)
مدل موجود	۰.۳۱۸۸	۱۵.۸۲	۱۲.۵۶
مدل رگرسیون با دو خروجی	۰.۱۹۴۴	۱۱.۱۳	۸.۷۹
مدل رگرسیون با دو ورودی	۰.۱۷۹۰	۱۱.۸۸	۸.۳۶
مدل سری زمانی(شبکه عصبی)	۰.۱۲۳۲	۷.۹۲	۴.۳۴

شکل (۳) نتیجه پیاده‌سازی مدل سری زمانی باشبکه عصبی است که از ۱۰ داده اول برای ساخت مدل استفاده شده است و مابقی نتیجه پیش‌بینی‌های صورت گرفته است. در این شکل به وضوح دیده می‌شود که پیش‌بینی شبکه عصبی بسیار به داده CDC نزدیک است.

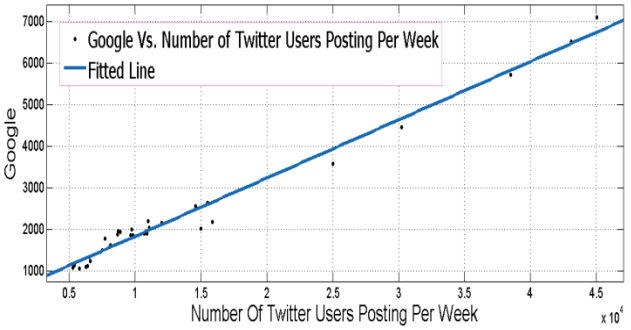


شکل(۳) : پیاده‌سازی مدل سری زمانی با شبکه عصبی

به‌دلیل اینکه CDC بین یک یا دو هفته طول می‌کشد که اطلاعات بیماری را منتشرکند، با پیش‌بینی توسط توییت‌های هر هفته می‌توان یک یا دو هفته جلوتر از این مرکز بود حال آنکه در مدل‌های پیشنهادی از توییت‌های هفته‌های قبل برای پیش‌بینی استفاده می‌گردد یعنی پیش‌بینی برای هفته بعد انجام می‌گیرد به‌عبارت دیگر دو یا چهار هفته از اعلام CDC زودتر پیش‌بینی صورت می‌گیرد.

۳-۳- پیاده‌سازی مدل برروی داده‌گوگل

گوگل نیز به پیش‌بینی روند شیوع آنفلوآنزا می‌پردازد. شکل (۴) ارتباط بین داده پیش‌بینی گوگل و داده توییتر را نشان می‌دهد. بنابراین برای اطمینان از درستی کار می توان بجای CDC این داده را مورد استفاده قرار داد. می‌خواهیم روش‌های پیاده‌سازی شده روی CDC را روی گوگل نیز انجام دهیم.



شکل(۴) : تعداد توییت‌های کاربران در هفته در مقابل داده گوگل

جدول (۳) نتیجه پیاده‌سازی روش‌های پیشنهادی را بر روی داده گوگل نشان می‌دهد. در اینجا نیز، در بین روش‌های ارائه شده مدل سری زمانی با شبکه عصبی جواب بهتری را می‌دهد.

جدول (۳) : نتایج پیاده‌سازی مدل‌ها روی داده گوگل

فرمول	RMSE	RMSE(درصد)	RAE(درصد)
مدل موجود	۴۰۳.۶۷۸۳	۲۲.۳۷	۲۲.۸۶
مدل رگرسیون با دو خروجی	۳۰۷.۶۲۸۷	۱۶.۰۳	۱۵.۶۸
مدل رگرسیون با دو ورودی	۲۴۹.۱۲۱۵	۱۳.۲۲	۱۲.۶۲
مدل سری زمانی(شبکه عصبی)	۸۸.۴۹۶۷	۴.۲۰	۴.۰۱

همان‌طور که مشاهده می‌شود، روش‌های ارائه شده بر روی داده گوگل نیز پاسخ مشابه با CDC و قابل قبولی را ارائه می‌دهد.

۴- خلاصه و نتیجه‌گیری

دراین مقاله، از روی داده‌های به دست آمده از توییتر به پیش‌بینی بیماری آنفلوآنزا پرداخته شده است. بدین منظور روش‌های رگرسیون خطی با ورودی خارجی و سری زمانی با شبکه عصبی را برای پیش-بینی درصد بیماران ملاقات شده توسط پزشک ساختیم. مدل‌های ارائه شده بر روی داده CDC و گوگل پیاده‌سازی شد و مشخص شد که خطای جذر میانگین مربع‌ها و خطای نسبی مطلق در روش‌های پیشنهادی هشت تا هجده درصد کمتر از روش قبلی موجود است.

مراجع

[1] JORDANS, F., "*WHO working on formulas to model swine flu spread"*, 2009, http://seattletimes.com/html/nationworld/2010823058_fluvaccine18.html.

[2] "Analysis of Spanish flu cases in 1918-1920 suggests transfusions might help in bird flu pandemic", http://www.eurekalert.org/pub_releases/2006-08/acop-aos082806.php

[3] Longini, I. M., Nizam, A., Xu, S., Ungchusak, K, Hanshaoworakul, W., Cummings, D. A. and Halloran, M. E., *"Containing pandemic influenza at the source"*, Science, Vol. 309, no. 5737, pp. 1083-1087, 2005.

[4] *"Centers for Disease Control and Prevention"*, a weekly influenza surveillance report, 2009, http://www.cdc.gov/.

[5] Magruder, S., "Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease", Johns Hopkins University APL Technical Digest, 2003.

[6] "Twitter," Wikipedia, http://en.wikipedia.org/wiki/Twitter.

[7] Culotta, A., "Towards detecting influenza epidemics by analyzing Twitter messages", Proceedings of the First Workshop on Social Media Analytics - SOMA '10, pp. 115–122, 2010.

[8] Bodnar, T. and Salathé, M., "Validating Models for Disease Detection Using Twitter Regression on Tweet Count", Proceedings of the 22nd international conference on World Wide Web companion, pp. 699–702, 2013.

[9] Sadilek, A., Kautz, H. and Silenzio, V., "Modeling Spread of Disease from Social Interactions", International AAAI Conference on Weblogs and Social Media, 2012.

[10] Peng, H.-K., Zhu, J., Piao, D., Yan, R. and Zhang, Y., "Retweet Modeling Using Conditional Random Fields", IEEE 11th International Conference on Data Mining Workshops, pp. 336–343, Dec 2011.

[11] Sadilek, A., Kautz, H. and Silenzio, V., "Predicting Disease Transmission from Geo-Tagged Micro-Blog Data", Twenty-Sixth AAAI Conference, 2012.

[12] *"UMass Lowell Wiki"*, http://91-541.wiki.uml.edu/SNEFT

[13] H. Achrekar, R. Lazarus, and W. C. Park, "Predicting Flu Trends using Twitter Data," IEEE Infocom, pp. 702–707, 2011.

[14] Motoyama, M., Voelker, G. M. and Savage, S., "Measuring Online Service Availability Using Twitter", WOSN'10 Proceedings of the 3rd conference on Online social networks, pp. 13, 2010.

[15] Mislove, A., "Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter", 2010, http://www.ccs.neu.edu/home/amislove/twittermood.

[16] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. and Brilliant, L., "Detecting influenza epidemics using search engine query data.", Nature, vol. 457, no. 7232, pp. 1012–1014, Feb. 2009.

[17] Lampos, V. and Cristianini, N., "Tracking the flu pandemic by monitoring the social web", 2nd International Workshop on Cognitive Information Processing, pp. 411–416, Jun 2010.

[18] Paul, M. J. and Dredze, M., "You Are What You Tweet: Analyzing Twitter for Public Health", AAAI Publications, Fifth International AAAI Conference on Weblogs and Social Media, 2011.

[19] Veloso, A. and Ferraz, F., "Dengue surveillance based on a computational model of spatio-temporal locality of Twitter", ACM WebSci'11, pp.14-17, 2011.

[20] *"SNEFT"*, http://www.cs.uml.edu/~hachreka/SNEFT

[21] Achrekar, H., Gandhe, A., Lazarus, R., Yu, S. and Liu, B., "Online Social Networks Flu Trend Tracker - A Novel Sensory Approach to Predict Flu Trends", Springer, vol. 357, pp. 353–368, 2013 .

[22] "Google Flu Trend", www.google.org/flutrends/, accessed on Nov. 11, 2013.

زیرنویس‌ها

[1] Centers for Disease Control and Prevention(CDC)
[2] Influenza Like Illness(ILI)
[3] Conditional Random Field(CRF)
[4] Social Network Enabled Flu Trends(SNEFT)
[5] retweet
[6] Syndrome Elapse Time(SET)
[7] Cross Validation(CV)
[8] Auto-regression with exogenous inputs(ARX)
[9] Root Mean Square Error(RMSE)
[10] Relative Absolute Error(RAE)
[11] Nonlinear autoregressive exogenous(NARX)