پیش بینی ورود گردشگران با شاخص یادگیری ماشین و جستجو در اینترنت

چکیده

مطالعات قبلی نشان داده اند که داده های آنلاین، مانند استفاده از موتورهای جستجو، یک منبع جدید از اطلاعات هستند که می توانند برای پیش بینی تقاضای گردشگری مورد استفاده قرار بگیرند. در این مطالعه، ما یک چارچوب برای پیش بینی پیشنهاد می کنیم که از شاخص های یادگیری ماشین و جستجو در اینترنت برای پیش بینی ورود گردشگران به مقصد های محبوب در چین استفاده کرده و عملکرد پیش بینی آن را با نتایج جستجوی بدست آمده از گوگل و بایدو مقایسه می کند. این تحقیق ارتباط Granger و هم جمعی بین شاخص جستجوی اینترنتی و ورود گردشگران پکن را بررسی می کند. نتایج تجربی ما نشان می دهد که در مقایسه با مدل های معیار، مدل های مبتنی بر ماشین یادگیری سریع با هسته غیر خطی پیشنهاد شده (KELM) که مجموعه های گردشگری را با شاخص بایدو و شاخص گوگل ادغام می کنند، می توانند عملکرد پیش بینی را به طور قابل توجهی از لحاظ دقت پیش بینی و آنالیز قدرت بهبود بخشد.

کلیدواژه ها: پیش بینی تقاضای گردشگری، ماشین یادگیری سریع با هسته غیر خطی، داده های حاصل از جستجو، آنالیز داده های بزرگ، شاخص جستجوی کامپوزیت

۱. مقدمه

در سراسر جهان، صنعت گردشگری به طور قابل توجهی به رشد اقتصادی کمک می کند (Gunter & Onder، 2015؛ Song، Li، Witt، Athanasopoulos & ، 2011). طبق گزارش اداره گردشگری ملی چین، درآمد گردشگری چین در سال ۲۰۱۶ به ۴,۶۹ تریلیون RMB رسید که این مقدار نسبت به سال قبل ۱۳,۶ درصد افزایش داشته است و ۶,۳ درصد از GDP (تولید ناخالص داخلی) چین را تشکیل می دهد. بنابراین پیش بینی میزان گردشگری به طور فزاینده ای برای پیش بینی توسعه اقتصادی آینده مهم است. پیش بینی تقاضای گردشگری ممکن است اطلاعات اولیه را برای برنامه ریزی و سیاست گذاری بعدی ارائه دهد (Chu، 2008؛ Witt & Song، 2002). روش

های مورد استفاده برای مدل سازی و پیش بینی گردشگری به چهار گروه مدل های سری زمانی، مدل های اقتصادسنجی، تکنیک های هوش مصنوعی و روش های کیفی تقسیم می شوند (Goh & Law، 2011؛ Song & Li، 2008). علاوه بر اطلاعات توریستی ساده ارائه شده توسط اداره آمار دولتی، جستجو در اینترنت که منعکس کننده رفتار و اهداف گردشگران است، به طور فزاینده در مدل های پیش بینی گردشگری مورد استفاده قرار گرفته است (Croce، 2017؛ Goodwin، 2008). با این حال، شاخص جستجو فرصت های زیادی را برای فرآیند مدل سازی پیش بینی گردشگری ایجاد کرده است (Li، Pan، Raw & Huang، 2017). داده های جستجو در اینترنت برای جنبه های زیادی مانند ثبت نام هتل (Pan & Yang، 2017، Rivera، 2016)، شماره های توریستی (Bangwayo-Skeete & Skeete 2015، Yang، Pan، Evans، Lv &، 2015)، شاخص های اقتصادی (Choi & Varian، 2012)، نرخ بیکاری (Askitas & Zimmermann، 2009)، هزینه های خصوصی (Vosen & Schmidt، 2011) و بازده سهام (Zhu & Bao، 2014) مورد استفاده قرار گرفته است. هنگام معرفی شاخص Baidu یا شاخص Google به مدل های پیش بینی، کلمات کلیدی و ترکیب شاخص ها باید با دقت انتخاب شوند. کلمات کلیدی را می توان با توجه به ضریب همبستگی، نمودار گرایش یا روش مربع جمعیت انتخاب کرد (Brynjolfsson، Geva، Reichman، 2016). علاوه بر این، ترکیب شاخص ها را می توان با استفاده از روش HE-TDC (Peng, Liu, Wang, & Gu, 2017) و یا آنالیز مولفه اصلی (PCA) بدست آورد. بدیهی است که باید تا حد امکان برای جلوگیری از مشکلات مربوط به چند خطی و بیش برازش تلاش کرد.

در این مطالعه، ما یک چارچوب جدید یکپارچه سازی یادگیری ماشین و شاخص جستجوی اینترنتی برای پیش بینی حجم گردشگری پیشنهاد کردیم. در این تحقیق، ما یک چارچوب جدید به منظور یکپارچه سازی یادگیری ماشین و شاخص جستجوی اینترنتی برای پیش بینی حجم گردشگری پیشنهاد کردیم. قدرت پیش بینی این چارچوب به دو ویژگی بستگی دارد: اول، جستجوی اینترنتی مربوطه به شدت به تناسب کمک می کند؛ دوم، ماشین های یادگیری مبتنی بر Kernel دارای زمان محاسبات کوتاه و توانایی تعمیم خوب هستند. با این حال، تا آنجا که ما می دانیم، تعداد کمی از مطالعات از ماشین یادگیری سریع برای پیش بینی تقاضای گردشگری استفاده کرده اند. چارچوب

پیشنهادی برای پیش بینی ورود گردشگران به پکن مورد استفاده قرار گرفته است. کلمات کلیدی مربوط به جستجوی اینترنتی، جنبه های مختلف گردشگری از جمله غذا، مسکن، تفریح، خرید، سفر و ترافیک را پوشش می دهد. بر خلاف مطالعات قبلی، این مقاله هر دو شاخص بایدو و شاخص گوگل را در بر می گیرد، که وضعیت فعلی گردشگران داخلی و مسافران خارجی را منعکس می کند. نتایج تجربی نشان می دهد که چارچوب پیش بینی پیشنهادی به طور قابل توجهی برتر از مدل های سنتی سری زمانی و برخی از دیگر مدل های یادگیری ماشین است. در عین حال، قدرت پیش بینی مدل ها با وجود شاخص بایدو و شاخص گوگل بسیار بیشتر از حالتی است که بدون یک شاخص و یا بدون هر دو شاخص باشد، که این ممکن است مدرکی برای این امر باشد که جستجوی اینترنتی اهمیت زیادی برای پیش بینی تقاضای گردشگری دارد.

ادامه این مقاله به شرح زیر است: بررسی ادبیات در بخش ۲ ارائه شده است. ماشین یادگیری سریع با هسته غیر خطی در بخش ۳ معرفی شده است. چارچوب پیش بینی در بخش ۴ نشان داده شده است. مطالعه تجربی در بخش ۵ داده شده است. در نهایت، بخش ۶ نتیجه کار و مفاهیم تحقیقات بیشتر را ارائه می دهد.

منابع	منطقه متمرکز شده	اهداف تحقیق	فرکانس داده ها	متدولوژی ها	اندازه گیری عملکرد	متغیرها
Athanasopoulos and Hyndman (۲۰۰۸) Bangwayo-Skeete & Skeete, 2015 Chen, Lai, and Yeh (۲۰۱۲) Chu (۲۰۰۸) Fildes, Wei, and Ismail (۲۰۱۱) Gunter and Onder (۲۰۱۵)	استرالیا کارائیب تایوان آسیایی-اقیانوس آرام انگلستان پاریس	گردشگری درون مرزی تقاضای گردشگری گردشگری درون مرزی تقاضای گردشگری تقاضای مسافرت هوایی	فصلی ماهیانه ماهیانه، ماهیانه، فصلی سالیانه ماهیانه	SSME, ESAR-MIDAS EMD, BPNN مدل ARAR ADLM, TVP, VAR EC-ADLM, VAR, Bayesian VAR, TVP,	RMSE, ME, MAE, MAPE MAPE, RMSE, DM MAD, MAPE, RMSE MAPE, RMSE MAE, RMSE	ورود گردشگران و متغیرهای اقتصادی ورود گردشگران، داده های گوگل ترند گردشگران ورودی گردشگران ورودی

		گردشگری درون مرزی		ARMA, ETS	RMSE, MAE	مسافران هوایی و متغیر اقتصادی گردشگران ورودی به هتل ها و متغیر اقتصادی
Jungmittag (۲۰۱۶) Li, Pan, Law, and Huang (۲۰۱۷) Liang (۲۰۱۴) Pan and Yang (۲۰۱۷)	آلمان پکن تایوان شهرستان چارلستون	تقاضای سفر تقاضای گردشگری گردشگری درون مرزی تقاضای هتل	ماهیانه ماهیانه ماهیانه هفتگی	پیش بینی های ترکیب، SARIMA GDFM, PCA SARIMA, GARCH ARIMAX	MAE, RMSE, MAPE MAE, MAPE MAD, RMSE, MAPE MAPE, RMSE	مسافران هوایی گردشگران ورودی و شاخص بایدو گردشگران ورودی سکونت در هتل، پرسش های موتور جستجو، وب سایت رفت و آمد، اطلاعات آب و هوا
Du Preez and Witt (۲۰۰۳) Rivera (۲۰۱۶)	سیشل پورتوریکو	گردشگری درون مرزی رزرو هتل	ماهیانه ماهیانه	ARIMA, SSM DLM	MAE, RMSE, MAPE MAE, MAPE, RMSE	گردشگران ورودی و متغیرهای اقتصادی داده های گوگل ترند، NHNR

Shahrabi, Hadavandi, and Asadi (۲۰۱۳)	ژاپن	گردشگری درون مرزی	ماهیانه	MGFFS	RMSE, MAPE	گردشگران ورودی
Song et al. (۲۰۱۱) Wong, Song, Witt, and Wu (۲۰۰۷)	هنگ کنگ هنگ کنگ	گردشگری درون مرزی گردشگری درون مرزی	ماهیانه فصلی	STSM, TVP، ARIMA، ADLM، ECM، VAR، ترکیب پیش بینی	MAPE, RMSE MAPE, RMSE, MAE	گردشگران ورودی و متغیرهای اقتصادی گردشگران ورودی و متغیرهای اقتصادی
Wu and Cao (۲۰۱۶)	سرزمین اصلی چین	گردشگری درون مرزی	ماهیانه	SVR, FOA, SIA	MAPE, RMSE, R	جریان گردشگری درون مرزی

جدول ۱. مرور اجمالی از مطالعات برگزیده پیش بینی توریست

۲. بررسی ادبیات

این بخش، مقالات و ادبیات مربوط به پیش بینی ورود گردشگران و پیش بینی گردشگری با داده های بدست آمده از موتور جستجو را بررسی می کند. لیستی از این ادبیات در جدول ۱ ارائه شده است.

نکات: فضای حالت مدلسازی شده با متغیرهای خارجی (SSME)؛ مدل هموارسازی نمایی (ES)؛ مدل های نمونه گیری داده های ترکیبی اتورگرسیو (خود برگشتی) (AR-MIDAS)؛ تجزیه حالت تجربی (EMD)؛ شبکه عصبی پس انتشار (BPNN)؛ مدل تاخیر توزیع شده اتورگرسیو (ADLM)؛ پارامتر زمان متغیر (TVP)؛ مدل اتورگرسیو برداری (VAR)؛ اصلاح خطای مدل تاخیر توزیع شده اتورگرسیو (EC-ADLM)؛ اتورگرسیو برداری بیزین (BVAR)؛ اتورگرسیو میانگین متحرک (ARMA)؛ میانگین متحرک اتورگرسیو یکپارچه فصلی (SARIMA)؛ مدل فاکتور حرکتی تعمیم یافته (GDFM)؛ آنالیز مولفه اصلی (PCA)؛ ناهم واریانسی مشروط اتورگرسیو عمومی (GARCH)؛ میانگین متحرک اتورگرسیو یکپارچه با متغیرهای خارجی (ARIMAX)؛ مدل های فضای حالت (SSM)؛ مدل خطی پویا (DLM)؛ سیستم فلزی ژنتیکی پیش بینی مدولار (MGFFS)؛ مدل سری زمانی ساختاری (STSM)؛ رگرسیون بردار پشتیبانی

(SVR)؛ الگوریتم بهینه سازی مگس میوه (FOA)؛ تنظیم شاخص فصلی (SIA)؛ خطای جذر میانگین مربعات (RMSE)؛ خطای میانگین (ME)؛ میانگین خطای مطلق (MAE)؛ میانگین خطای درصد مطلق (MAPE)؛ آمار Diebold-Mariano (DM)؛ میانگین انحراف مطلق (MAD)؛ تعداد ثبت نام های غیرمقیم هتل (NHNR).

۲،۱ پیش بینی میزان گردشگری

میانگین متحرک اتورگرسیو یکپارچه (ARIMA) بیشترین مدل پیش بینی سری زمانی است که مورد استفاده قرار گرفته است. این مدل همچنین به طور گسترده برای پیش بینی گردشگری و اجرای خوب آن مورد استفاده قرار گرفته است (Athanasopoulos, Hyndman, Song, & Wu, 2011; Brida & Risso, 2011; Chang & Liao, 2010; Chen et al., 2012; Du Preez & Witt, 2003; Jungmittag, 2016; Li & Sheng, 2016; Liang, 2014; Lim & McAleer, 2002; Shahrabi et al., 2013). با این حال، مدل های ARIMA همیشه برتر از دیگر مدل ها نیستند. این مدل ها در مدل های اقتصادسنجی سنتی به خوبی عمل می کنند، اما گاهی اوقات در روش های هوش عملکرد ضعیفی دارند. به همین ترتیب، تحقیقات قبلی در مورد ورود گردشگران به چین نشان داد که رگرسیون بردار پشتیبانی (SVR) عملکرد بهتری از شبکه عصبی پس انتشار (BPNN) و مدل های ARIMA داشته است.

هموارسازی نمایی (ES) به طور گسترده در پیش بینی گردشگری مورد استفاده قرار گرفته است و بسیاری از محققان از این مدل به عنوان یک معیار استفاده می کنند (Fildes et al.، 2011؛ Park، Rilett & Han، 1999؛ Witt & Witt، 1995). اما دیگر مدل های سری زمانی ذکر شده از جمله مدل های اصلاح خطا (ECM) (Lee, 2011; Shen, Li, & Song, 2009; Vanegas, 2013; Wong et al., 2007)، و مدل های ناهم واریانسی مشروط اتورگرسیو عمومی (GARCH) (Chan، Lim، Mcaleer & 2005؛ Liang، 2014)، محدود به مدل های فضای حالت نیستند (Athanasopoulos & Hyndman، 2008؛ Beneki، Eecels & Leon، ۲۰۱۲؛ Du Preez & Witt، 2003).

در سال های اخیر، تکنیک های هوش مصنوعی (AI) در مطالعات گردشگری از قبیل نظریه منطق فازی، شبکه های عصبی مصنوعی (ANNs)، ماشین های بردار پشتیبانی (SVM) و الگوریتم های ژنتیکی (GA) پدیدار شده اند. مزیت اصلی و کلیدی AI این است که آنها به هیچ یک از فرض هایی مانند مانایی یا توزیع نیاز ندارند. از این رو، این تکنیک های AI به طور گسترده برای پیش بینی تقاضای گردشگری مورد استفاده قرار گرفته اند.

به عنوان مثال، ANN ها تکنیک های محاسباتی ساده ای هستند که در علوم کامپیوتر و سایر رشته های تحقیقاتی مورد استفاده قرار می گیرند. ویژگی های منحصر به فرد ANNها از جمله سازگاری و غیر خطی بودن، این تکنیک را به یک جایگزین مناسب برای مدل های پیش بینی رگرسیون تبدیل می کند (Song & Li، 2008). اولین مدل محاسباتی برای ANNها توسط McCulloch و Pitts (۱۹۴۳) با استفاده از الگوریتم های منطقی آستانی و ریاضیات پیشنهاد شده است. علاوه بر این، ANN دارای یک توسعه سریع و بهبود مستمر است. این نشان داده شده است که عملکرد کلی ANNها برای پیش بینی گردشگری، بهتر از مدل های سنتی سری زمانی و مدل های اقتصادسنجی است. استفاده از ANN برای پیش بینی ورود گردشگران ژاپنی در هنگ کنگ به مراتب ساده تر بوده و عملکرد بهتری در رگرسیون های متعدد، هموارسازی نمایی و میانگین متحرک دارد (Law & Au، 1999).

به طور کلی SVMها برای حل طبقه بندی، تخمین رگرسیون و پیش بینی مشکلات مورد استفاده قرار می گیرند. SVM در ابتدا در اواخر دهه ۱۹۹۰ برای پیش بینی گردشگری معرفی شد و نسخه های اصلاح شده آن همچنان تا بعد از سال ۲۰۰۰ معرفی می شدند. Sencheong و Turner (۲۰۰۵) به بررسی ادبیات مربوط به کاربردهای SVM برای پیش بینی گردشگری پرداختند. نتایج تجربی نشان می دهد که SVM معمولا عملکرد بهتری نسبت به مدل های سری زمانی و مدل های رگرسیون چندگانه در پیش بینی گردشگری ارائه می دهد. به عنوان مثال، Claveria، Monte و Torra (۲۰۱۶) نشان دادند که SVM عملکرد پیش بینی را با توجه به مدل معیار بهبود داد. نتایج مشابه بدست آمده توسط Pai، Hung و Lin (۲۰۱۴) نیز نشان می دهد که عملکرد SVM در پیش بینی تقاضای گردشگری در هنگ کنگ و تایوان بهتر از مدل ARIMA است.

۲,۲ پیش بینی گردشگری با استفاده از داده های موتور جستجو

در سال های اخیر تعدادی از نویسندگان توجه بیشتری به کاربرد داده های جستجو در وب برای پیش بینی گردشگری داشته اند. یک جستجوی رایج در Google Trends یا Baidu Index، که از عمده کلمات کلیدی برای جستجوی اینترنتی هستند، می تواند برای شناسایی گردشگران بالقوه و همچنین به عنوان یک شاخص از رفتارهای گردشگری، از جمله مکان و چگونگی سفر گردشگران، مورد استفاده قرار گیرد.

استفاده از داده های جستجو در وب می تواند به طور قابل توجهی دقت پیش بینی حجم گردشگری را بهبود بخشد. به عنوان مثال، Pan، Wu و Song (۲۰۱۲) نشان دادند که زمانی که داده های جستجوی Google در مدل ARMA گنجانده شده باشد، دقت پیش بینی به طور قابل توجهی بهبود می یابد که این امر یک پشتیبانی قوی برای استفاده از داده های موتورهای جستجو در پیش بینی تقاضای اتاق های هتل ارائه می دهد. نتایج مشابهی توسط Artola، Pinto و Garcia (۲۰۱۵) به دست آمده است، که پیش بینی های ورود گردشگران به اسپانیا را با استفاده از شاخص های گوگل در اندازه گیری میزان جستجوهای اینترنتی در رابطه با کلمات کلیدی مربوط به سفر به اسپانیا را بهبود می بخشد. علاوه بر این، شاخص بایدو برای پیش بینی تقاضای گردشگری در چین مورد استفاده قرار گرفته است. Yang و همکاران (۲۰۱۵) از داده های موتور جستجو برای پیش بینی ورود گردشگران به ایالت هینان استفاده کردند. نتایج تجربی آنها نشان داد که داده های بدست آمده از موتور جستجوی گوگل و بایدو عملکرد پیش بینی را به طور قابل توجهی بهبود بخشید؛ علاوه بر این، داده های شاخص بایدو به دلیل سهم بیشتر آن در بازار چین عملکرد ارائه داده اند. به همین ترتیب، Li و همکاران (۲۰۱۷) از داده های موتور جستجو برای ایجاد یک شاخص جستجوی کامپوزیتی و از یک مدل فاکتور حرکتی تعمیم یافته (GDFM) برای پیش بینی ورود گردشگران پکن استفاده کردند.

۳. ماشین یادگیری سریع با هسته غیر خطی

دستگاه یادگیری افراطی (ELM) نوعی از شبکه های عصبی پیش خور با یک تک لایه ی پنهان (SLFN ها) است (Huang، Zhu، Siew &، 2006). مدل ELM به واسطه سرعت یادگیری سریع و توانایی تعمیم خود به طور گسترده در بسیاری از زمینه ها مورد استفاده قرار گرفته است. ویژگی منحصر به فرد مدل ELM این است که وزن های ورودی

و بایاس ها به طور تصادفی تولید شده و پارامترهای لایه پنهان نیاز به تنظیم ندارند. وزن های خروجی با استفاده از محاسبات ماتریسی ساده به دست می آیند، بنابراین زمان محاسبات بسیار کوتاه است.

برای N نمونه دلخواه (x_i, y_i)، $x_i \in R^N$، $y_i \in R^N$، $i = 1,2,\cdots,N$، اگر تابع فعال سازی لایه پنهان $h(x)$ و ماتریس خروجی Y باشد، SLFNهای معمول را می توان به صورت زیر تعریف کرد:

$$Y = \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{mj} \end{bmatrix}_{m\times N} = \begin{bmatrix} \sum_{i=1}^{l} \beta_{i1} h(w_i x_j + b_i) \\ \sum_{i=1}^{l} \beta_{i2} h(w_i x_j + b_i) \\ \vdots \\ \sum_{i=1}^{l} \beta_{im} h(w_i x_j + b_i) \end{bmatrix}_{m\times N}, (j = 1,2, \cdots N) \qquad (1)$$

که در آن β نشان دهنده وزن های خروجی شبکه بین لایه پنهان و لایه خروجی، $W_i=[\omega_{i1},\omega_{i2},\ldots,\omega_{iN}]^T$ نشان دهنده وزن ورودی بین iامین لایه پنهان و لایه های ورودی، l تعداد گره های پنهان و b آستانه لایه پنهان است. معادلات فوق را می توان به صورت زیر هم نوشت:

$$H\beta = Y, \quad Y \in \Re^{N\times m}, \beta \in \Re^{N\times m} \quad , \quad H = H(\omega, b) = h(\omega x + b) \qquad (2)$$

که در آن H ماتریس خروجی لایه پنهان است. وزن ها و بایاس های لایه ورودی به جای تنظیم، بطور تصادفی بدست آمده اند. تنها پارامتر مجهول وزن خروجی β است که می توان آن را با روش معمولی کمترین مربعات (OLS) حل کرد. راه حل معادله فوق به شکل زیر داده شده است:

$$\hat{\beta} = H^{\dagger}Y, \quad H^{\dagger} = H^T(HH^T)^{-1} \qquad (3)$$

که در آن $H^{\dagger}$ نشان دهنده معکوس تعمیم یافته H Moore-Penrose است (Huang et al., 2006). با توجه به تئوری رگرسیون ریج و روش تصویر متعامد، β را می توان با اضافه کردن یک فاکتور مثبت 1/C به شکل زیر محاسبه کرد:

$$\hat{\beta} = H^T(1/C + HH^T)^{-1}Y \qquad (4)$$

سپس، تابع خروجی ELM را می توان به شرح زیر بیان کرد:

$$f(x) = H\hat{\beta} = HH^T(1/C + HH^T)^{-1}Y \quad (5)$$

این روش بعضی از ضعف های الگوریتم های یادگیری مبتنی بر گرادیانت معمولی از قبیل بیش برازش، کمینگی محلی و زمان محاسبه طولانی را برطرف می کند. ساختار توپولوژی (مکان شناسی) ELM در شکل ۱ نشان داده شده است. یک ELM مبتنی بر هسته غیر خطی توسط Huang (۲۰۱۴) پیشنهاد شد. در نظریه هوانگ، تابع فعال سازی h(x) لایه پنهان با یک تابع هسته ای در شرایط Mercer جایگزین شده است. تابع خروجی KELM را می توان به صورت زیر فرموله کرد:

$$f(x) = h(x)\hat{\beta} = \begin{bmatrix} k(x, x_1) \\ k(x, x_2) \\ \vdots \\ k(x, x_n) \end{bmatrix}^T (1/C + HH^T)^{-1}Y \quad (6)$$

در این فرمول، نیازی نیست که نگاشت ویژگی h(x) برای کاربران مشخص باشد؛ به جای آن می توان از هسته منتاظر با آن $k(x, x_i)$ استفاده کرد. این وضعیت بدان معنی است که یک تابع هسته ای می تواند نگاشت تصادفی ELM را جایگزین کند تا وزن خروجی پایدارتر باشد. بنابراین، توانایی تعمیم KELM بهتر از ELM است. در این مقاله، چهار تابع هسته ای مختلف به شرح زیر مورد استفاده قرار گرفته است:

(۱) تابع هسته ای خطی:

$$K(x, x_i) = x^T x_i \quad (7)$$

(۲) تابع هسته ای چند جمله ای:

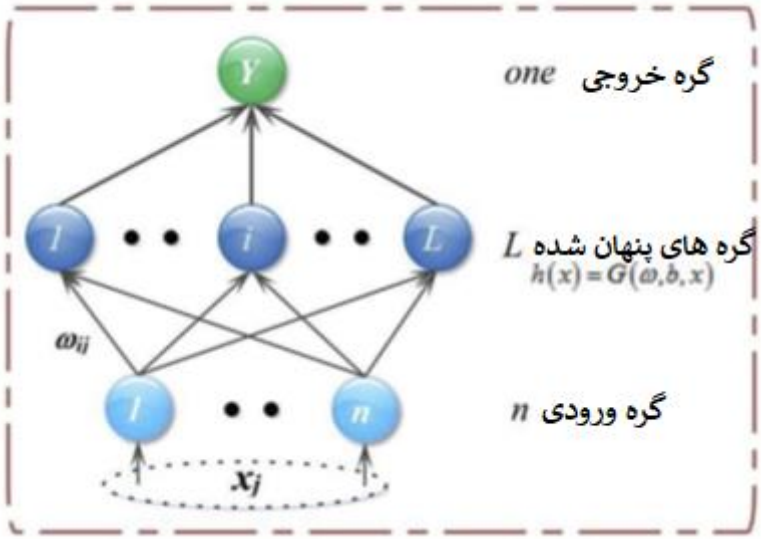
$$K(x, x_i) = (\gamma x^T x_i + r)^p, \quad \gamma > 0 \quad (8)$$

(۳) تابع هسته ای RBF:

$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2), \quad \gamma > 0 \quad (9)$$

(۴) تابع هسته ای موجک:

$$K(x, x_i) = \cos(\alpha*(x - x_i))\exp(-\gamma(x - x_i)^2), \quad \alpha, \gamma > 0 \qquad (10)$$



شکل ۱. ساختار توپولوژی ELM

۴. چارچوب پیش بینی

در طول روند برنامه ریزی برای سفر، بازدیدکنندگان باید تصمیمات زیادی در مورد تمام جنبه های سفر مانند انتخاب مقصد، ترافیک، محل اسکان، و غذا اتخاذ کنند. پیش از ورود، بازدیدکنندگان باید این تصمیمات را براساس زمان خود بگیرند که این امر بین افراد مختلف متفاوت است. بازدیدکنندگان اغلب برای کمک در تصمیم گیری از موتورهای جستجو استفاده می کنند. بنابراین، ممکن است انواع مختلف اطلاعات مورد نیاز بازدیدکنندگان در زمان های مختلف با استفاده از این موتورهای جستجو بدست آیند. از این رو، ما یک چارچوب پیش بینی ارائه می دهیم که از ماشین های یادگیری و شاخص های جستجوی اینترنتی برای پیش بینی ورود گردشگران استفاده می کند (شکل ۲). این چارچوب شروع فرآیند مدل سازی را از طریق استخراج، ادغام و محاسبات داده ها توصیف می کند.

۵. مطالعه تجربی

پکن به عنوان محل مورد نظر برای این مطالعه تجربی به منظور ارزیابی اثربخشی چارچوب پیش بینی پیشنهادی انتخاب شد. طراحی تجربی در بخش ۵,۱ معرفی شده و نتایج تجربی در بخش ۵,۲ ارائه شده است. در نهایت این نتایج در بخش ۵,۳ خلاصه شده و مورد بحث قرار می گیرند.

شکل ۲. چارچوب پیش بینی با یادگیری ماشینی و شاخص جست و جوی اینترنت

۵,۱ طراحی تجربی

۵,۱,۱ جمع آوری داده ها

داده های مربوط به ورود ماهانه گردشگران به پکن که برابر مجموع ورود گردشگران داخلی و خارجی است، در طی دوره ژانویه ۲۰۱۱ تا آوریل ۲۰۱۷ از پایگاه داده Wind تهیه شده است (http://www.wind.com.cn/). داده ها به زیر مجموعه های نمونه و زیر مجموعه های غیر نمونه تقسیم شدند. همانطور که در شکل ۳ نشان داده شده است، زیر مجموعه های نمونه برای پرورش مدل با داده های ژانویه ۲۰۱۱ تا آوریل ۲۰۱۶ مورد استفاده قرار گرفته است، در حالی که زیر مجموعه های غیر نمونه برای آزمایش تجربی با داده های مه ۲۰۱۶ تا آوریل ۲۰۱۷ به کار گرفته شده است. داده های دقیق را می توان از پایگاه داده Wind یا با درخواست از نویسندگان دریافت کرد.

علاوه بر این، به منظور بررسی نتایج جستجوی اینترنتی در موتورهای جستجوی مختلف برای پیش بینی ورود گردشگران، این مطالعه داده های حاصل از جستجو در دو موتور جستجو اصلی Baidu Index (http://index.baidu.com/) و Google Trends (https://trends.google.com/trends/) را جمع آوری کرد. بایدو تقریبا ۸۰,۵٪ سهم بازار در چین را به خود اختصاص داده و برای نمایش روش جستجوی گردشگران داخلی در این مقاله مورد استفاده قرار گرفته است؛ گوگل محبوب ترین موتور جستجوی جهان است که حدود ۹۲,۵٪ از بازار را تشکیل می دهد و برای نشان دادن روش جستجوی گردشگران خارجی در این مقاله به کار گرفته شده است. تاریخچه جستجوی ایجاد شده توسط دو موتور جستجو در دسترس عموم است. اگرچه Google Trends و Index Baidu شاخص های خود را با روش های مختلفی محاسبه می کنند، اما هر دو نشان دهنده محبوبیت جستجوی خاص و منافع کاربر در زمان های خاص هستند (Yang et al., 2015). بنابراین، به منظور مقایسه این دو موتور جستجو، ما داده های جستجوی ماهانه هر دو موتور جستجو را به ترتیب از ژانویه ۲۰۱۱ تا آوریل ۲۰۱۶ انتخاب کردیم. بخش های بعدی یک روش سیستماتیک برای انتخاب کلمات کلیدی جستجو و ساخت شاخص های جستجوی اینترنتی برای پیش بینی ورود گردشگران به پکن را با جزئیات کامل شرح می دهند.

۵,۱,۲ معیارهای ارزیابی

به منظور بررسی دقت پیش بینی مدل های مختلف، ما دو معیار ارزیابی اصلی را برای مقایسه عملکرد پیش بینی زیر مجموعه های نمونه و غیر نمونه اتخاذ کردیم: خطای جذر میانگین مربعات نرمال (NORMSE) و میانگین خطای درصد مطلق (MAPE).

$$NRMSE = \frac{100}{\bar{x}}\sqrt{\frac{1}{N}\sum_{t=1}^{N}(x_t - \hat{x}_t)^2}\,,\ MAPE = \frac{100}{N}\sum_{t=1}^{N}\left|\frac{x_t - \hat{x}_t}{x_t}\right|, \qquad (11)$$

که در آن N تعداد مشاهدات، x_t نشان دهنده حجم گردشگری واقعی و $\hat{x}_t$ نشان دهنده مقدار پیش بینی شده برای حجم گردشگری است.

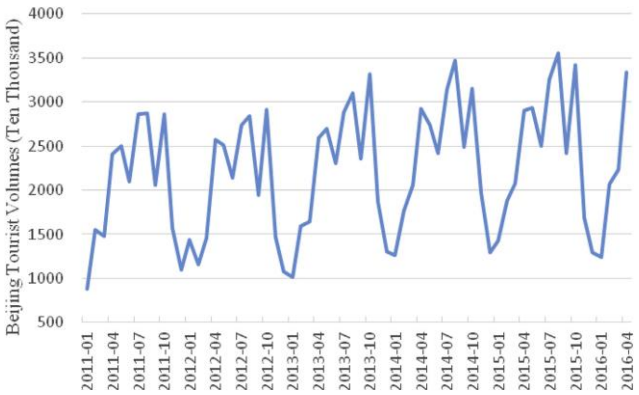
علاوه بر این، به منظور ارزیابی عملکرد پیش بینی از یک دیدگاه آماری، Diebold-Mariano (DM) آماری برای تست اهمیت آماری تمام مدل ها مورد استفاده قرار گرفت (Diebold & Mariano، 2002). از DM آماری برای تست فرضیه پوچ برابری دقت پیش بینی پیش مورد نظر در مقابل جایگزینی توانایی های مختلف پیش بینی در مدل ها استفاده شد. در این مطالعه، از خطای پیش بینی میانگین مربعات (MSPE) به عنوان عملکرد از دست رفته استفاده شد. بنابراین، فرضیه پوچ تست DM این بود که MSPE مدل تست شده te، کوچکتر از مدل مدل معیار be نیست. برای می توان DM آماری برای مدل تست شده te و مدل معیار be را به صورت زیر تعریف کرد:

$$S_{DM} = \frac{\bar{g}}{\left(\widehat{V}_{\bar{g}}/N\right)^{1/2}} \qquad (12)$$

که در آن $\bar{g} = \left(\sum_{t=1}^{N} g_t\right)/N \;\; \left(g_t = \sum_{t=1}^{N} (x_t - \hat{x}_{te,t})^2 - \sum_{t=1}^{N} (x_t - \hat{x}_{be,t})^2\right)$ و $\widehat{V}_{\bar{g}} = \gamma_0 + 2\sum_{l-1}^{\infty} \gamma_l \; (\gamma_l = \mathrm{cov}(g_t, g_{t-l}))$. $\hat{x}_{te,t}$ and $\hat{x}_{be,t}$ مقادیر پیش بینی x_t هستند که به ترتیب از طریق مدل تست شده te و مدل معیار be برای مدت زمان t محاسبه شده اند.

۵,۲ نتایج تجربی

در این بخش، ما ابتدا دو شاخص جستجوی اینترنتی بایدو و گوگل را با استفاده از شاخص جستجوی پیشرو کامپوزیت ایجاد می کنیم. در مرحله دوم، ما آزمون هم جمعی و علیت گرنجر را بین ورود گردشگران و این دو شاخص انجام می دهیم. در مرحله سوم، ما مدل های پیش بینی مختلفی از لحاظ سری گردشگران ورودی و این دو شاخص طراحی کرده و عملکرد پیش بینی زیر مجموعه های نمونه و غیر نمونه در هر مدل را ارزیابی می کنیم. در نهایت، قدرت تمام مدل های پیش بینی شده آنالیز شده و نتیجه گیری نهایی از طریق این آنالیز تجربی صورت می گیرد.



شکل ۳. روند ماهانه توریست های وارد شده به پکن

۵,۲,۱ شاخص جستجو در اینترنت

این زیربخش یک فرآیند پنج مرحله ای برای انتخاب جستجوهای کاندید و ساخت یک شاخص جستجوی اینترنتی با استفاده از کارکرد جستجوهای مرتبط با Baidu Index و Google Trends را نشان می دهد (Yang et al., 2015):

(۱) ما در ابتدا ۲۴ پرسش اصلی جستجو شده را بر اساس تمام جنبه های برنامه ریزی گردشگری از جمله سفر، ترافیک، محل اسکان، غذاخوری، تفریح و خرید انتخاب کردیم (Li et al., 2017). پرسش ها با توجه به گروه مربوطه در جدول ۲ ذکر شده اند.

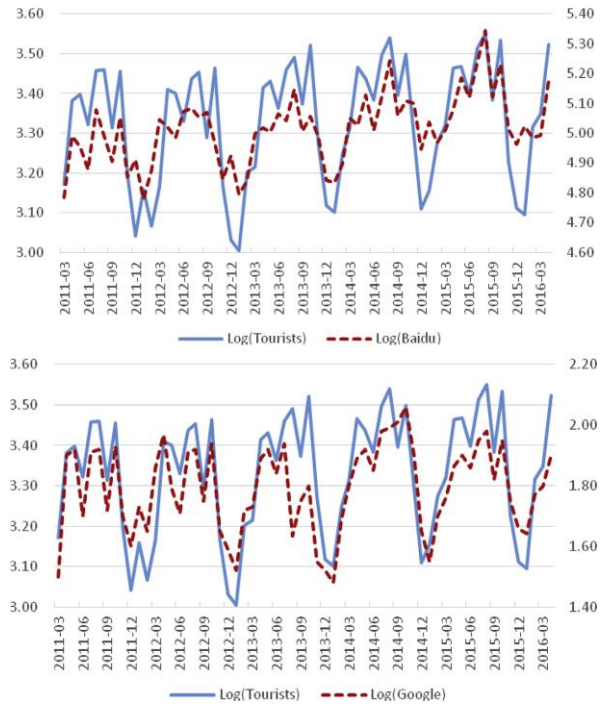
(۲) ما ابتدا این ۲۴ کلیدواژه را در Baidu Index و Google Trends جستجو کرده و کلیدواژه های مرتبط بازیابی کردیم. سپس ما دوباره کلید واژه های معرفی شده را به عنوان دور دوم کلید واژه ها در نظر می گیریم. این فرآیند برای چندین دور تکرار شد و در طی آن کلیدواژه های با داده های غیر قابل دسترس یا داده های بسیار کم حذف شدند. تعداد کلیدواژه ها به ۱۵۴ عدد برای Baidu Index و ۶۹ عدد برای Google Trends رسید.

(۳) ما ضریب همبستگی پیرسون را با دوره های تاخیر مختلف بین گردشگران ورودی پکن و هر یک از کلمات کلیدی جستجو محاسبه کردیم. مجموع چهار ضریب همبستگی برای هر کلمه کلیدی جستجو، شامل ارتباط بین ورودی های گردشگری از دوره جاری و حجم کلمات جستجو شده به ترتیب در ۰ تا ۳ ماه گذشته محاسبه شد. علاوه بر این، ما در هنگام ساخت شاخص جستجو در اینترنت خود، کلمات کلیدی با بالاترین ضریب همبستگی را انتخاب کردیم. مجموعه

ای از ۲۴ کلید واژه از Baidu Index و ۱۶ کلید واژه از Google Trends انتخاب شده است (در جدول ۳ و جدول ۴ نشان داده شده است). برای بدست آوردن تعداد واقعی کلید واژه ها، ما از یک ضریب همبستگی آستانه بین گردشگران ورودی و دو شاخص موتورهای جستجو استفاده کردیم: ۰,۷۵ برای بایدو و ۰,۷ برای گوگل.

(۴) برای پیش بینی ورود گردشگران در آینده، ما تنها کلید واژه هایی را انتخاب کردیم که حداقل یک دوره تاخیر قبل از ماه ورود گردشگران داشته باشند، زیرا بایدو و گوگل تنها در پایان هر ماه اطلاعات را منتشر می کنند. در نهایت، مجموعه ای از ۲۲ کلید واژه با یک یا دو دوره تاخیر به عنوان پیش بینی کننده Baidu Index و ۱۳ کلید واژه به عنوان پیش بینی کننده Google Trends انتخاب شد.

(۵) ما از طریق جمع و جابجایی، اطلاعات بیشتری را برای یک شاخص کامپوزیت جمع کردیم. تمام کلید واژه های انتخاب شده از طریق تاخیر حداکثر ضریب همبستگی پیرسون منتقل شده و تمام کلید واژه های جستجوی منتقل شده در همان مدل، برای تشکیل یک سری زمانی جدید جمع شدند. شکل ۴ همبستگی بین ورود ماهانه گردشگران به پکن و دو شاخص جستجو در اینترنت را نشان می دهد. آنالیزهای بعدی مبتنی بر شاخص جستجوی اینترنت کامپوزیت برای هر دو داده های بایدو و گوگل بودند.



شکل ۴. روند ماهانه توریست های وارد شده به پکن و دو شاخص جست و جوی اینترنت

پرسش های جستجو شده		پرسش های جستجو شده		پرسش های جستجو شده	
اسکان		رفت و آمد		گردشگری	
9	هتل های پکن	5	خطوط هوایی پکن	1	گردشگری پکن
10	اقامتگاه های پکن	6	اتوبوس شاتل پکن	2	آب و هوای پکن
11	خانه های روستایی پکن	7	راه آهن پکن	3	نقشه های پکن
12	استراحتگاه های پکن	8	بلیط ها اتوبوس های پکن	4	سفر به پکن
	خرید		برنامه زمانی		آژانس
21	مراکز خرید پکن	17	تفریحات		غذا
22	کالاهای خاص پکن	18	تفریحات شبانه پکن	13	غذاهای پکن
23	خیابان داشیلان	19	امکانات تفریحی پکن	14	مرغابی پکن
24	مرکز خرید Panjiayuan	20	کافه های پکن نمایش های پکن	15	وب سایت های غذای پکن
				16	تنقلات پکن

جدول ۲. پرسشنامه جست و جوی مرتبط با صنعت توریسم پکن

No.	پرسش های جستجو شده	ترتیب تاخیر	پرسش های جستجو شده	ترتیب تاخیر
1	آژانس مسافرتی پکن	2	مرکز ملی آبی	1
2	راه حل سفر به پکن	2	ورزشگاه ملی	1
3	تنقلات پکن	1	برج تلویزیونی مرکزی	1
4	هتل های پکن	1	دیوار بزرگ بادلینگ	1
5	خانه های روستایی پکن	1	رزرو هتل	1
6	راهنمایی اقامتگاه های پکن	1	فرودگاه های پکن	1
7	گردشگری پکن	1	پرواز های پکن	1
8	راهنمایی سفر به پکن	1	پارک های تفریحی پکن	1
9	سایت های سفر به پکن	1	مرکزPanjiayuan	1
10	موزه کاخ	1	خیابان داشیلان	1
11	منطقه شیدان	1	آب و هوای پکن	0
12	مقبره های مینگ	1	نقشه های پکن	0

جدول ۳. حداکثر ضرایب همبستگی پرسشنامه جست و جوی بایدو

N	پرسش های جستجو شده	Lag order	N	پرسش های جستجو شده	ترتیب تاخیر
1	سفر به چین	2	9	سفر به پکن	1
2	آب و هوای پکن	2	10	دیوار بزرگ	1
3	مرغابی پکن	1	11	پرواز های پکن	1
4	دستور پخت مرغابی	1	12	فرودگاه های پکن	1
5	هتل های پکن	1	13	خطوط راه آهن پکن	1
6		1	14	نقشه های پکن	0
7	رستوران های پکن	1	15	کافه های پکن	0
8	مراکز خرید پکن	1	16	نمایش های پکن	0
	Zhongguanc				

جدول ۴. حداکثر ضرایب همبستگی پرسشنامه جست و جوی گوگل

۵,۲,۲ آنالیز همبستگی و علیت گرنجر

به منظور کاهش تأثیر بخش های مجزا، این سه متغیر به فرم لگاریتمی تبدیل شدند (LogT، LogBI و LogGI). جدول ۵ آزمون پایداری و آزمون هم جمعی جوهانسون بین LogT، LogBI و LogGI را ارائه می دهد. این سه سری زمانی با استفاده از آزمون Dicky-Fuler ثابت در نظر گرفته شده اند. نتایج آزمون هم جمعی نشان می دهد که LogT و LogBI یکپارچه هستند. به همین ترتیب، LogT و LogGI نیز همپوشانی دارند. بنابراین، یک رابطه بلندمدت هم جمعی بین شاخص های جستجو در اینترنت و ورود گردشگران به پکن وجود دارد. این یافته ها نشان می دهد که اتخاذ شاخص های جستجو در اینترنت برای پیش بینی ورود گردشگران، از دیدگاه اقتصاد سنجی امکان پذیر است.

هدف از آزمون علیت گرنجر شناسایی این است که آیا این دو شاخص جستجو در اینترنت پیش بینی کننده ورود گردشگران پکن هستند یا خیر. همانطور که در جدول ۶ نشان داده شده است، LogGI و LogGI علت گرنجر LogT هستند، که این امر نشان دهنده ارتباط علی بین داده های این دو شاخص جستجوی اینترنتی و گردشگران ورودی واقعی پکن است.

تست های Dickey-Fuler تکمیل شده

	آمار t	مقدار p
Log*T*	−3.8056	0.0027
Log*BI*	−3.7143	0.0041
Log*GI*	−3.4034	0.0115

هم جمعی بین LogT و LogBI

	مقدار خاص	ردیابی آمار	مقدار بحرانی	Prob[b]
هیچ e[a]	0.07	27.31	15.98	0.00
حداکثر ۱ 1[a]	0.05	14.11	4.02	0.00

هم جمعی بین LogT و LogGI

	مقدار خاص	ردیابی آمار	مقدار بحرانی	Prob[b]
None[a]	0.06	17.58	15.43	0.01
At most 1[a]	0.03	1.19	3.42	0.22

آ) علامت گذاری رد فرضیه باطل در سطح اطمینان ۰،۰۵

ب) مککینون، هاگ و میشلیس (۱۹۹۹) p-value

جدول ۵. نتایج تست هم جمعی

۵،۲،۳ پیش بینی با یادگیری ماشین و شاخص جستجو در اینترنت

تکنیک های یادگیری ماشین برای بررسی بیشتر قدرت پیش بینی شاخص های جستجوی اینترنتی برای پیش بینی ورود گردشگران به پکن مورد استفاده قرار گرفته است. متغیرهای مستقل مدل های پیش بینی به چهار دسته تقسیم شدند: "سری زمانی"، "سری زمانی + شاخص بایدو"، "سری زمان + شاخص گوگل"، و "سری زمان + شاخص بایدو + شاخص گوگل". در این مطالعه، ورودی های مدل های یادگیری ماشین و تعداد نورون های پنهان مدل های ANN و KELM با استفاده از آزمون و خطا برای به حداقل رساندن خطاهای پیش بینی نمونه گیری تعیین شدند. تابع هسته ای گاوس در مدل های LSSVR و SVR اعمال شدند. شکل مطلوب مدل ARIMA با به حداقل رساندن معیار شوارتز (SC) و معیار اطلاعات Akaike (AIC) برآورد شد.

عملکرد پیش بینی شده چهار متغیر مستقل مختلف و هشت مدلی که در بالا ذکر شد، در این بخش ارائه شده است. جدول ۷ نتایج مقایسه ای معیارهای ارزیابی MAPE و NRMSE را نشان می دهد.

Nul	فرضیه پوچ	آمار F	Prob.
Log	علت گرنجر LogT نیست LogBI	27.86	0.00[a]
Log	علت گرنجر LogBI نیست LogT	0.43	0.53
Log	علت گرنجر LogT نیست LogGI	26.17	0.00[a]
L	علت گرنجر LogGI نیست LogT	0.02	0.94

آ) سطح اهمیت ۱٪ را نشان می دهد

جدول ۶. آزمون علیت گرنجر میان شاخص جست و جو در اینترنت و ورود توریست

همانطور که جدول ۷ نشان می دهد، مدل پیشنهادی KELM-rbf با متغیرهای مستقل "سری زمانی + شاخص گوگل + شاخص بایدو" دارای کمترین MAPE و NRMSE در پیش بینی نمونه است. علاوه بر این، در پیش بینی خارج از نمونه، "سری زمانی + شاخص بایدو + شاخص گوگل" به طور مداوم عملکرد بهتری نسبت به دیگر متغیرهای مستقل در پیش بینی ورود گردشگران پکن در رابطه با MAPE و NRMSE دارد، بعد از آن "سری زمانی + شاخص بایدو" و "سری زمانی + شاخص گوگل"، و در نهایت "سری زمانی"در آخرین رتبه قرار دارد. علاوه بر این، مدل های KELM پیشنهادی، به ترتیب ۴۱,۳-۸,۵۳٪ MAPE کوچکتر و ۹,۳۲-۴,۱۵٪ NRMSE کوچکتر از مدل های ARIMAX را به ترتیب با نرخ دقت ۰,۶۴۳٪ و ۰,۷۲٪ در نمونه خارجی ایجاد کردند.

مدل ها	درون نمونه		خارج از نمونه	
	MAPE (%)	NRMSE (%)	MAPE (%)	NRMSE (%)
Time series				
ARIMA	8.142	9.061	9.168	10.021
ANN	3.071	3.689	4.067	4.967
SVR	2.953	3.267	3.591	4.301
LSSVR	2.916	3.261	3.407	4.016
KELM-lin	2.637	3.014	3.056	3.986
KELM-poly	1.784	2.569*	1.921	2.914
KELM-rbf	1.627*	2.571	1.709*	2.726*
KELM-wav	1.709	2.602	1.846	2.843
Time series + Baidu Index				
ARIMAX	5.516	5.763	5.962	6.167
ANN	2.571	2.698	2.423	2.564
SVR	1.914	2.069	2.196	2.306
LSSVR	1.706	1.817	1.834	1.902
KELM-lin	1.047	1.156	1.314	1.397
KELM-poly	0.972	1.098	1.196	1.264
KELM-rbf	0.958*	1.006*	1.026*	1.127*
KELM-wav	0.969	1.037	1.088	1.191
Time series + Google Index				
ARIMAX	5.967	6.129	6.118	6.237
ANN	2.261	2.342	2.367	2.468
SVR	2.174	2.228	2.306	2.325
LSSVR	1.918	2.106	2.118	2.267
KELM-lin	1.446	1.609	1.546	1.674
KELM-poly	1.369	1.438	1.438	1.598
KELM-rbf	1.297	1.392	1.357	1.416
KELM-wav	1.011*	1.126*	1.348*	1.425*
Time Series + Baidu Index + Google Index				
ARIMAX	4.593	4.672	4.054	4.856
ANN	1.698	1.783	1.967	2.016
SVR	1.732	1.914	1.933	2.065
LSSVR	1.426	1.678	1.704	1.816
KELM-lin	0.814	0.973	0.896	1.013
KELM-poly	0.674	0.784	0.792	0.804
KELM-rbf	0.492*	0.622*	0.643*	0.702*
KELM-wav	0.571	0.713	0.725	0.891

شماره های ستاره دار کمترین میزان خطا را نشان می دهد (MAPE و NRMSE).

جدول ۷. ارزیابی عملکرد پیش بینی

۵،۲،۴ آزمون DM برای پیش بینی خارج از نمونه

برای ارزیابی دقت پیش بینی مدل های مختلف از یک دیدگاه آماری، ما آزمون DM را به هشت مدل با چهار متغیر مستقل مختلف اعمال کردیم. نتایج آزمون DM در جدول ۸ نشان داده شده است. هنگامی که سری زمانی، شاخص بایدو و شاخص گوگل یکپارچه شدند، آمار DM و مقادیر p برای مدل KELM-rbf به ترتیب کمتر از ۶،۱۱۲۵- و

تقریبا صفر بود. این نشان می دهد که عملکرد مدل KELM-rbf به طور قابل توجهی بهتر از مدل های معیار دیگر با سطح اطمینان ۱۰۰٪ است. این آنالیز ها نتایج جالبی نشان دادند: (۱) هنگامی که مدل های KELM به عنوان هدف آزمون در نظر گرفته شدند، تمام مقادیر p کمتر از ۰,۰۰ بود که این امر نشان می دهد که مدل های KELM به طور قابل توجهی برتر از دیگر مدل های معیار و با سطح اطمینان تقریبا ۱۰۰٪ هستند؛ (۲) عملکرد پیش بینی SVR و ANN کاملا مشابه بوده و هیچ یک از آنها از لحاظ آماری برتری نسبت به دیگری نداشت. (۳) مدل ARIMAX دارای ضعیف ترین عملکرد پیش بینی با چهار متغیر مستقل متفاوت است.

۵,۲,۵ آنالیز قدرت

قدرت هشت مدل پیش بینی با چهار متغیر مستقل مختلف در این بخش ارزیابی شده است. از آن جایی که مدل های ARIMAX، ANN، SVR، LSSVR و KELM تمایل دارند که نتایج مختلف پیش بینی را با تنظیمات اولیه مختلف ایجاد کنند، ما تمام مدل های پیش بینی را بیست بار اجرا کرده و قدرت آنها را با توجه به انحراف معیار MAPE و NRMSE مورد آنالیز قرار دادیم. این تحلیل ها در جدول ۹ ارائه شده و شواهدی ارائه می دهند که (۱) KELM پایدار ترین مدل در میان تمام مدل های پیش بینی است، زیرا انحراف استاندارد آن از NRMSE و MAPE بسیار کوچکتر از سایر مدل های معیار است؛ (۲) تمام مدل های پیش بینی مبتنی بر "سری زمانی + شاخص بایدو + شاخص گوگل" قوی ترین رویکرد را دارند و انحراف استاندارد آنها از MAPE و NRMSE بسیار کمتر از همه مدل های متناظر است؛ (۳) ARIMAX در بین تمام مدل های پیش بینی با متغیرهای مستقل متفاوت، ناپایدارترین است.

۵,۳ خلاصه

برای خلاصه کردن:

(۱) عملکرد پیش بینی "سری زمانی + شاخص بایدو + شاخص گوگل" برتر از سایر متغیرهای مستقل است و پس از آن "سری زمانی + شاخص بایدو" و "سری زمانی + شاخص گوگل"، و در آخر "سری زمانی" در آخرین رتبه قرار دارند.

(۲) با توجه به پیچیدگی ذاتی داده های ورود گردشگران، تکنیک های AI بسیار مناسب تر از مدل ARIMAX در پیش بینی ورود گردشگران هستند.

(۳) عملکرد مدل های KELM پیشنهادی با توابع هسته های مختلف، بهتر از همه مدل های معیار دیگر در پیش بینی دقت و قدرت است.

(۴) با توجه به آنالیز دقت و قدرت، قدرت پیش بینی مدل های پیشنهادی KELM پس از LSSVR، SVR، ANN و ARIMAX پایدارترین و موثرترین هستند.

۶. نتیجه گیری

در این مقاله، ما یک چارچوب پیش بینی پیشنهاد کردیم که از یادگیری ماشین و شاخص های جستجو در اینترنت برای پیش بینی ورود گردشگران به مقصد های محبوب در چین استفاده کرده و عملکرد پیش بینی آن را برای داده های جستجوی ایجاد شده توسط گوگل و بایدو را مقایسه می کند. این تحقیق ارتباط هم جمعی و رابطه علیت گرنجر را بین شاخص جستجوی اینترنتی و حجم گردشگران در پکن مورد بررسی قرار داد. نتایج تجربی نشان می دهد که مدل های KELM پیشنهادی با مجموعه حجم گردشگری یکپارچه شاخص بایدو و شاخص گوگل می توانند عملکرد پیش بینی را به طور قابل توجهی بهبود بخشند. در مقایسه با سایر روش های پیش بینی معیارهای محبوب، مدل KELM ما، که "مجموعه حجم گردشگری + شاخص بایدو + شاخص گوگل" را ادغام می کند، دقیق تر و قوی تر است. در نتیجه، مدل KELM ما یک رویکرد امیدوار کننده برای حل مشکلات در پیش بینی حجم جریان های گردشگری است.

مطالعه ما می تواند تا حدودی الهام بخش باشد. اول، پیش بینی حجم گردشگری به دقت می تواند به کارکنان گردشگری به منظور بهینه سازی منابع، تخصیص و فرموله کردن راهبردهای قیمت گذاری به صورت منطقی کمک کند. علاوه بر این، پیش بینی حجم گردشگری به طور دقیق ممکن است به صنایع مختلفی که به طور مستقیم یا غیر مستقیم به گردشگری بستگی دارد، کمک کند. دوم اینکه، این شواهد محکمی به سیاست گذاران ارائه داده و روندهای حجم گردشگری را پیش بینی می کند که می تواند برای تنظیم تصمیمات سیاسی، طراحی زیرساخت ها برای برنامه

ریزی مسکونی گردشگری و سیستم حمل و نقل به دولت کمک کند. سوم، داده های بدست آمده از موتور جستجو یک داده تولید شده توسط کاربر است و می تواند به طور رایگان از موتور جستجوی اصلی به دست آورد. بسیاری از مطالعات تایید کرده اند که این می تواند به طور قابل توجهی دقت پیش بینی ورود گردشگران را بهبود بخشد. بنابراین، این می تواند به عنوان یک منبع داده جایگزین برای مدیریت منابع گردشگری مورد استفاده قرار گیرد. از این رو، مدیران گردشگری می توانند چارچوب پیش بینی پیشنهادی ما را برای پیش بینی تقاضای گردشگری از طریق جمع آوری داده های جستجوی اینترنتی مختلف برای تصمیم گیری استراتژیک در کاربرد عملی به کار ببرند.

علاوه بر پیش بینی جریان حجم گردشگری، چارچوب پیش بینی پیشنهادی با استفاده از یادگیری ماشین و شاخص جستجو در اینترنت می تواند برای حل دیگر مشکلات پیچیده و دشوار پیش بینی، از جمله پیش بینی روند سهام، پیش بینی قیمت نفت خام، و پیش بینی نرخ ارز مورد استفاده قرار گیرد.

با این حال، این مطالعه محدودیت هایی دارد که عمدتا به این دلیل است که ما فقط از سفر پکن به عنوان یک مورد آزمون استفاده کردیم. توانایی خلاصه کردن راه انتخاب کلیدواژه، و همچنین مفاهیمی که در آن کلیدواژه ها بدون توجه به آنچه که موتور جستجوگر انتخاب می کنند، همگرا خواهند بود، محدود است. تحقیقات بیشتر برای بررسی استفاده از داده های جستجو در اینترنت در سایر اهداف و همچنین تحقیقات تجربی با یک نمونه بزرگتر برای رسیدگی به این محدودیت ها ضروری است. علاوه بر این، با توجه به تغییرات مداوم در نیازهای اطلاعاتی کاربر وب، ایجاد یک راه پویا و جامع برای انتخاب کلمات کلیدی که می تواند به طور موثر با رقابت متغیر در بازار مقابله کند، باید مسیر تحقیق در آینده باشد.

تضاد منافع

نویسندگان اعلام می کنند که هیچ تضادی در منافع مربوط به انتشار این مقاله وجود ندارد.

فهرست مشارکت نویسندگان

Shaolong Sun و Yunjie Wei ایده ارائه شده را درک کردند. Shaolong Sun چارچوب پیش بینی را توسعه داده و محاسبات را انجام داد. Shaolong Sun و Yunjie Wei در تفسیر نتایج کمک کردند.Shouyang Wang

و Kwok-Leung Tsui، Shaolong Sun و Yunjie Wei را تشویق کردند تا داده های جستجوی اینترنتی را بررسی کرده و بر نتایج این تحقیق نظارت کنند. Shaolong Sun کار نوشتن دست نوشته را رهبری کرد. همه نویسندگان یک بازخورد انتقادی ارائه داده و به تحقیق، آنالیز و نوشتن دست نوشته کمک کردند. همه نویسندگان دست نوشته را خوانده و تایید کردند.

مدل تست شده	مدل مرجع			
	LSSVR	SVR	ANN	ARIMAX
Time series				
KELM-rbf	−6.2687 (0.0000)	−8.6782 (0.0000)	−11.6592 (0.0000)	−15.2679 (0.0000)
LSSVR		−0.6874 (0.2459)	−1.5789 (0.0572)	−8.2698 (0.0000)
SVR			−0.9876 (0.1617)	−8.0694 (0.0000)
ANN				−7.1627 (0.0000)
Time series + Baidu Index				
KELM-rbf	−5.9372 (0.0000)	−8.1695 (0.0000)	−10.0128 (0.0000)	−14.6872 (0.0000)
LSSVR		−1.7929 (0.0365)	−2.6197 (0.0044)	−6.0158 (0.0000)
SVR			−0.1185 (0.4528)	−5.0246 (0.0000)
ANN				−4.7691 (0.0000)
Time series + Google Index				
KELM-wav	−5.8564 (0.0000)	−7.9854 (0.0000)	−9.8756 (0.0000)	−14.6872 (0.0000)
LSSVR		−1.7543 (0.0397)	−2.5876 (0.0048)	−6.1229 (0.0000)
SVR			−0.2069 (0.4180)	−5.1267 (0.0000)
ANN				−4.5968 (0.0000)
Time Series + Baidu Index + Google Index				
KELM-rbf	−6.1125 (0.0000)	−8.5692 (0.0000)	−9.6485 (0.0000)	−14.1167 (0.0000)
LSSVR		−1.8467 (0.0324)	−2.5691 (0.0051)	−5.0168 (0.0000)
SVR			−0.2182 (0.4136)	−4.9613 (0.0000)
ANN				−4.6916 (0.0000)

جدول ۸. نتایج تست DM در مجموع داده های نمونه

Std.[a]	مدل های پیش بینی							
	ARIMAX	ANN	SVR	LSSVR	KELM-lin	KELM-poly	KELM-rbf	KELM-wav
	Time series							
Std. of MAPE	0.0029	0.0032	0.0028	0.0033	0.0021	0.0003	0.0001	0.0005
Std. of NRMSE	0.0081	0.1012	0.0084	0.0088	0.0067	0.0045	0.0039	0.0042
	Time series + Baidu Index							
Std. of MAPE	0.0031	0.0029	0.0030	0.0032	0.0014	0.0001	0.0000	0.0001
Std. of NRMSE	0.0078	0.0094	0.0079	0.0083	0.0052	0.0041	0.0023	0.0038
	Time series + Google Index							
Std. of MAPE	0.0027	0.0026	0.0023	0.0033	0.0018	0.0000	0.0000	0.0001
Std. of NRMSE	0.0072	0.0095	0.0073	0.0072	0.0041	0.0036	0.0023	0.0033
	Time Series + Baidu Index + Google Index							
Std. of MAPE	0.0025	0.0021	0.0018	0.0019	0.0016	0.0001	0.0001	0.0003
Std. of NRMSE	0.0071	0.0091	0.0062	0.0067	0.0035	0.0029	0.0018	0.0035

جدول ۹. تجزیه و تحلیل مقاومت

References

Artola, C., Pinto, F., & Garcia, P. D. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower, 36*, 103–116.

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly, 55*, 107–120.

Athanasopoulos, G., & Hyndman, R. J. (2008). Modelling and forecasting Australian domestic tourism. *Tourism Management, 29*, 19–31.

Athanasopoulos, G., Hyndman, R. J., Song, H. Y., & Wu, D. C. (2011). The tourism forecasting competition. *International Journal of Forecasting, 27*(3), 822–844.

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management, 46*, 454–464.

Beneki, C., Eeckels, B., & Leon, C. (2012). Signal extraction and forecasting of the UK tourism income time series: A singular spectrum analysis approach. *Journal of Forecasting, 31*, 391–400.

Brida, J. G., & Risso, W. A. (2011). Research note: Tourism demand forecasting with SARIMA models - the case of South Tyrol. *Tourism Economics, 17*, 209–221.

Brynjolfsson, E., Geva, T., & Reichman, S. (2016). Crowd-squared: Amplifying the predictive power of search trend data. *MIS Quarterly, 40*(4), 941–961.

Chang, Y. W., & Liao, M. Y. (2010). A seasonal ARIMA model of tourism forecasting: The case of Taiwan. *Asia Pacific Journal of Tourism Research, 15*, 215–221.

Chan, F., Lim, C., & Mcaleer, M. (2005). Modelling multivariate international tourism demand and volatility. *Tourism Management, 26*, 459–471.

Chen, C. F., Lai, M. C., & Yeh, C. C. (2012). Forecasting tourism demand based on empirical mode decomposition and neural network. *Knowledge-Based Systems, 26*, 281–287.

Choi, H. Y., & Varian, H. (2012). Predicting the present with Google trends. *The Economic Record, 88*, 2–9.

Chu, F. L. (2008). Forecasting tourism demand with ARMA-based methods. *Tourism Management, 30*, 740–751.

Claveria, O., Monte, E., & Torra, S. (2016). Combination forecasts of tourism demand with machine learning models. *Applied Economics Letters, 23*, 428–431.

Croce, V. (2017). Business confidence and international tourism Demand: Evidence from a global panel of experts. *Global Journal of Management and Business Research, 16*(1), 29–42.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 20*(1), 134–144.

Du Preez, J., & Witt, S. F. (2003). Univariate versus multivariate time series forecasting: An application to international tourism demand. *International Journal of Forecasting, 19*(3), 435–451.

Fildes, R., Wei, Y., & Ismail, S. (2011). Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting, 27*(3), 902–922.

Goh, C., & Law, R. (2011). The methodological progress of tourism demand forecasting: A review of related literature. *Journal of Travel & Tourism Marketing, 28*(3), 296–317.

Goodwin, P. (2008). A quick tour of tourism forecasting. *Foresight, 10*, 35–37.

Gunter, U., & Onder, I. (2015). Forecasting international city tourism demand for Paris: Accuracy of uni- and multivariate models employing monthly data. *Tourism Management, 46*, 123–135.

Huang, G. B. (2014). An insight into extreme learning machines: Random neurons, random features and kernels. *Cognitive Computation, 6*(3), 376–390.

Huang, G. B., Zhu, Q. Y., & Siew, C. K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing, 70*(1), 489–501.

Jungmittag, A. (2016). Combination of forecasts across estimation windows: An application to air travel demand. *Journal of Forecasting, 35*(4), 373–380.

Law, R., & Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management, 20*, 89–97.

Lee, K. N. (2011). Forecasting long-haul tourism demand for Hong Kong using error correction models. *Applied Economics, 43*, 527–549.

Liang, Y. H. (2014). Forecasting models for Taiwanese tourism demand after allowance for Mainland China tourists visiting Taiwan. *Computers & Industrial Engineering, 74*(1), 111–119.

Lim, C., & McAleer, M. (2002). Time series forecasts of international travel demand for Australia. *Tourism Management, 23*, 389–396.

Li, X., Pan, B., Law, R., & Huang, X. K. (2017). Forecasting tourism demand with composite search index. *Tourism Management, 59*, 57–66.

Li, Z. C., & Sheng, D. (2016). Forecasting passenger travel demand for air and high-speed rail integration service: A case study of Beijing-guangzhou corridor, China. *Transportation Research Part A: Policy and Practice, 94*(1), 397–410.

MacKinnon, J. G., Haug, A. A., & Michelis, L. (1999). Numerical distribution functions of likelihood ratio tests for cointegration. *Journal of Applied Econometrics, 14*, 563–577.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115–133.

Pai, P. F., Hung, K. C., & Lin, K. P. (2014). Tourism demand forecasting using novel hybrid system. *Expert Systems with Applications, 41*(8), 3691–3702.

Pan, B., Wu, C. D., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology, 3*(3), 196–210.

Pan, B., & Yang, Y. (2017). Forecasting destination weekly hotel occupancy with big data. *Journal of Travel Research, 56*(7), 957–970.

Park, D., Rilett, L. R., & Han, G. (1999). Spectral basis neural networks for real-time travel time forecasting. *Journal of Transportation Engineering, 125*(6), 515–523.

Peng, G., Liu, Y., Wang, J., & Gu, J. (2017). Analysis of the prediction capability of web search data based on the HE-TDC method-prediction of the volume of daily tourism visitors. *Journal of Systems Science and Systems Engineering, 26*(2), 163–182.

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google trends data. *Tourism Management, 57*, 12–20.

Sencheong, K., & Turner, L. W. (2005). Neural network forecasting of tourism demand. *Tourism Economics, 11*, 301–328.

Shahrabi, J., Hadavandi, E., & Asadi, S. (2013). Developing a hybrid intelligent model for forecasting problems: Case study of tourism demand time series. *Knowledge-Based Systems, 43*, 112–122.

Shen, S., Li, G., & Song, H. (2009). Effect of seasonality treatment on the forecasting performance of tourism demand models. *Tourism Economics, 15*(4), 693–708.

Song, H., & Li, G. (2008). Tourism demand modelling and forecasting-A review of recent research. *Tourism Management, 29*(2), 203–220.

Song, H. Y., Li, G., Witt, S. F., & Athanasopoulos, G. (2011). Forecasting tourist arrivals using time-varying parameter structural time series models. *International Journal of Forecasting, 27*, 855–869.

Song, H., & Witt, S. F. (2000). Tourism demand modelling and forecasting: Modern econometric approaches. *Journal of Retailing and Consumer Services, 9*(1), 54–55.

Vanegas, M. (2013). Co-integration and error correction estimation to forecast tourism in El Salvador. *Journal of Travel & Tourism Marketing, 30*, 523–537.

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting, 30*(6), 565–578.

Witt, S., & Song, H. (2002). Forecasting future tourism flows. In S. Medlik, & A. Lockwood (Eds.). *Tourism and hospitality in the 21st century* (pp. 106–118). Oxford: Butterworth-Heinemann.

Witt, S. F., & Witt, C. A. (1995). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting, 11*(3), 447–475.

Wong, K. K., Song, H., Witt, S. F., & Wu, D. C. (2007). Tourism forecasting: To combine or not to combine? *Tourism Management, 28*(4), 1068–1078.

Wu, L. J., & Cao, G. H. (2016). Seasonal SVR with FOA algorithm for single-step and multi-step ahead forecasting in monthly inbound tourist flow. *Knowledge-Based Systems, 110*, 157–166.

Yang, X., Pan, B., Evans, J. A., & Lv, B. F. (2015). Forecasting Chinese tourist volume with search engine data. *Tourism Management, 46*, 386–397.

Zhu, Y., & Bao, W. B. (2014). The impact of investors' attention on stock returns -study based on Baidu index. *Service systems and service management (ICSSSM), 2014 11th international conference on* (pp. 1–5). IEEE.