*چشم‌انداز مدیریت صنعتی*
شماره ۹ ـ بهار ۱۳۹۲
ص‌ص ۱۷۶ – ۱۵۹

# طراحی مدل پیش بینی تقاضا در صنعت کاشی و سرامیک

**علیرضا موتمنی*، مصطفی رضائی**، مریم احقاقی*****

## چکیده

پیش‌بینی تقاضا جزء مهم‌ترین فعالیتهای هر سازمان برای برنامه‌ریزی فروش و در نهایت برنامه ریزی جامع بوده و در واقع تعیین کننده حجم فعالیتهای سازمان در آینده می باشد. همچنین درک درستی از میزان وکیفیت فعالیتهای مزبور را برای مدیران فراهم می نماید. در این مقاله بر اساس مبانی علوم هوش‌مصنوعی و داده‌کاوی، به ارائه مدلی جهت پیش‌بینی میزان فروش در صنعت کاشی و سرامیک پرداخته شده است. مدل پیشنهادی، یک مدل ترکیبی شامل کاهش بعد، خوشه‌بندی و پیش‌بینی بوده و به منظور اجرای مراحل مختلف آن از الگوریتم‌های آنالیز مؤلفه‌های مستقل، یادگیری منیفلد، خوشه‌بندی کامینز و رگرسیون بردار پشتیبان، استفاده شده است. در تحقیق حاضر ۵۰ مورد از فروش های ماهیانه مربوط به ۳ سال گذشته شرکت کاشی ایرانا مورد استفاده قرار گرفته است. نتایج بدست آمده از طریق پیش بینی فروش با مدل پیشنهادی، بدلیل کاهش خطاهای عمومی و خطاهای نمونه در مقایسه با مدلهای سنتی پیش بینی، از دقت بالاتری برخوردار می باشد.

**کلیدواژه‌ها: پیشبینی تقاضا؛ رگرسیون بردار پشتیبان؛ یادگیری منیفلد؛ آنالیز مؤلفه‌های مستقل؛ خوشه‌بندی کامینز.**

---

تاریخ دریافت مقاله: ۹۱/۱۲/۲۴، تاریخ پذیرش مقاله: ۹۲/۴/۳۱.

* استادیار، دانشگاه شهید بهشتی (نویسنده مسئول).

E-mail: ar_motameni@yahoo.com

** کارشناسی ارشد، دانشگاه سمنان.

*** کارشناسی ارشد، دانشگاه تربیت مدرس.

**۱. مقدمه**

یکی از مسائل چالش‌برانگیز در بهبود عملکرد سازمان‌ها، پیش بینی تقاضا، بهبود زنجیره تأمین آن‌ها و کاهش هزینه‌های مرتبط با آن است. بنا بر آمار منتشر شده، در سال ۱۹۹۸ شرکت‌های آمریکایی، ۸۹۸ میلیارد دلار صرف هزینه‌های مرتبط با زنجیره تأمین کرده‌اند(این مبلغ با ۱۰ درصد تولید ناخالص ملی آن کشور در آن سال برابر است). این عدد به یک ترلیون دلار در سال ۲۰۰۰ افزایش یافت و در سال‌های ۲۰۰۱ و ۲۰۰۲ برابر با ۹۵۷ میلیارد دلار و ۹۱۰ میلیارد دلار گزارش شده است [۱]. متأسفانه بیشتر این هزینه‌ها غیرضروری بوده و با بهبود مدیریت زنجیره تأمین می‌توان به میزان قابل توجهی هزینه‌های مزبور را کاهش داد [۲:۷]؛ برای نمونه متخصصین معتقدند که با بهره‌گیری از استراتژی‌های بهتر زنجیره تأمین، صنایع غذایی می‌توانند حدود ۳۰ میلیارد دلار یا ۱۰ درصد از هزینه‌های عملیاتی سالانه خود بکاهند [۳]. از جمله راه‌حل‌های کاهش هزینه‌های زنجیره تأمین، افزایش دقت پیش‌بینی تقاضا به کمک روش‌های جدید آماری است. روش‌های سنتی پیش‌بینی تقاضا علاوه بر دقت پایین، مشکلات دیگری نیز برای سازمان ها به وجود می‌آورند؛ از جمله این مشکلات، می توان از«اثر شلاقی» یاد کرد [۲:۲۳]. با پیشرفت‌های اخیر در هوش مصنوعی، تکنیک‌های جدیدی برای پیش‌بینی ارائه شده است که نسبت به تکنیک‌های سنتی، از دقت بالاتری برخوردار هستند. رایج‌ترین این تکنیک‌ها، الگوریتم‌های شبکه عصبی است که کاستی‌هایی مانند نیاز به پارامترهای کنترلی زیاد، دشواری رسیدن به نتیجه‌ای پایدار و خطر برازش بیش از حد، دارد [۴، ۵، ۶]. به دلیل وجود چنین ضعف‌هایی، مدل‌های بهتری برای بهبود مدل شبکه عصبی طراحی شده است. ماشین‌های بردار پشتیبان از الگوریتم شبکه عصبی بدیعی بر مبنای تئوری یادگیری آماری بهره می‌گیرند [۷، ۸]. این الگوریتم‌ها پتانسیل بسیار بالایی داشته و در مسائل کاربردی نتایج بسیار خوبی به دنبال دارند. امروزه مدل رگرسیون ماشین‌های بردار پشتیبان به نام «رگرسیون بردار پشتیبان » نیز در موضوع پیش‌بینی‌های غیرخطی، بسیار مورد توجه قرار گرفته و نتایج بسیار خوبی به همراه داشته است. با این حال بسیاری از محققان یادآور این نکته شده‌اند که وجود اغتشاش در داده‌ها، نتایج پیش‌بینی را تا حد بسیار زیادی تحت تأثیر قرار می‌دهد [۹، ۱۰]. به نظر می‌رسد پیش‌پردازش داده‌ها بیش از هر زمانی اهمیت پیدا کرده است [۱۱].

با توجه به اهمیت موضوع پیش‌بینی تقاضا در سازمان‌ها، در این مقاله مدل سه‌مرحله‌ای برای پیش‌بینی فروش ارائه شده است به این صورت که قبل از انجام پیش بینی بوسیله رگرسیون بردار پشتیبان از دو مرحله خاص برای پیش بینی پردازش داده ها استفاده گردیده است.

در مرحله اول سعی می شود که با کاهش بعد داده‌ها اغتشاشات موجود در آن‌ها تا حد امکان گرفته شود و، در مرحله دوم با دسته‌بندی و خوشه‌بندی کردن داده‌ها و قرار دادن داده‌های مشابه

در گروه‌های یکسان، دقت پیش‌بینی افزایش داده شده و در نهایت پیش‌بینـی را روی داده‌هـایی که اغتشاشات آنها تا حد امکان گرفته شده انجام می‌گیرد.

## ۲. پیشینه تحقیق

در انجام این تحقیق از مفاهیم آنالیز اجزای اصلی، یادگیری منیفلد، آنالیز اجزای مستقل، خوشه‌بندی کامینز، ماشین بردار پشتیبان، رگرسیون بردار پشتیبان و بیش‌برازش استفاده شده که در ادامه، تعریف مختصری از مفاهیم مزبور و نتایج تحقیقات مرتبط ارائه گردیده است.

**آنالیز اجزای اصلی**[1]. آنالیز اجزای اصلی از روش‌های کلاسیک در آنالیز آماری داده[2]، استخراج ویژگی[3] و فشرده‌سازی داده به‌شمار می‌رود که از لحاظ تاریخی، به کارهای اولیه پیرسون[4] در حدود سال های ۱۹۰۰ بازمی‌گردد.

هدف اصلی آنالیز اجزای اصلی، پیدا کردن راسـتاهای مؤلفـه‌هـای اصـلی اسـت. راسـتاهای مؤلفه‌های اصلی راستاهایی هستند که تصویر داده‌ها بر آن‌ها حداکثر واریانس را خواهنـد داشـت؛ به این ترتیب، مؤلفه اصلی اول، راستایی است که تصویر داده‌ها بر آن حـداکثر واریـانس را دارد، دومین مؤلفه اصلی، راستایی است که از میان تمامی راسـتاهای عمـود بـر راسـتای اول، تصـویر داده‌ها ماکزیمم واریانس را دارا بوده و k امین مؤلفه‌ی اصلی، راستایی است کـه از میـان تمـامی راستاهای عمود بر k-1 راستای قبل، تصویر داده‌ها بر آن بیشترین واریانس را دارد.

برای پیدا کردن راستای همه مؤلفه‌های اصلی، محاسبات به صورت جبر خطی انجام خواهـد شد. اگر تمامی n داده مورد مطالعه را روی هم بگذاریم و در ماتریس X که یـک مـاتریس $n^{\times}p$ است، بنویسیم، تصویر داده‌ها Xw خواهد بود که یک مـاتریس $n^{\times}1$ بـوده و واریـانس آن برابـر خواهد بود با:

رابطه ۱
$$\sigma_{\vec{w}}^{2} = \frac{1}{n}\sum_{i} (\vec{x}_i \cdot \vec{w})^{2}$$

1. Principle Component Analysis
2. Statistical Data Analysis
3. Feature Extraction
4. Pearson

$$= \frac{1}{n}(Xw)^T (Xw)$$
$$= \frac{1}{n} w^T X^T X w$$
$$= w^T \frac{X^T X}{n} w$$
$$= w^T V w$$

می‌توان نشان دهیم که بردار w مورد نظر، بردار ویژه ماتریس کوواریانس V است. از آنجا که V یک ماتریس P×P است، تعداد P بردار ویژه متفاوت خواهد داشت.

**یادگیری منیفلد.** منیفلد یک فضای فرعی منحنی و چنبری است که در درون یک فضای اقلیدسی، جاسازی شده است.

نکته مهم در یادگیری منیفلد این است که هر منیفلد q بعدی را می‌توان با یک فضای فرعی خطی q بعدی تخمین زده و در اطراف هر نقطه فضای کوچکی را در نظر گرفته و فضای مماس بر آن را پیدا نمود. همچنین فضای مماس نقاط در طول منیفلد به صورت پیوسته تغییر می‌کند. هر چه تغییرات فضای مماسی بیشتر باشد، انحنای منیفلد نیز بیشتر خواهد بود. بنابراین اگر داده‌های ما روی یک منیفلد قرار گرفته باشد، می‌توانیم تخمین خطی از منیفلد به صورت محلی داشته باشیم.

بایستی توجه داشت که در یادگیری منیفلد برای کاهش بعد غیر خطی از روشهای مختلفی استفاده می شود. یکی از روش‌های کاهش بعد غیرخطی الگوریتم LLE[1] است که به اختصار در زمینه روش مزبور توضیح داده می شود.

الگوریتم LLE سه مرحله دارد؛ در مرحله اول برای هریک از نقاط یک همسایگی تشکیل داده می شود، در مرحله دوم برای تخمین خطی نقاط بر اساس همسایگی آن‌ها اقدام گردیده و محاسبه وزن‌ها تخمین زده شده و در مرحله سوم بر اساس وزن‌های پیداشده، مختصات منیفلد پیدا می گردد.

مراحل الگوریتم LLE به صورت زیر است:

۱. برای هر داده ، k نزدیک‌ترین همسایه شاخص می شود.

1. Locally Linear Embedding

۲. ماتریس وزن w به صورتی پیدا می گردد کـه جمـع مجـذور خطاهـای بازسـازی داده هـا از همسایه‌هایش حداقل شود:

رابطه ۲

$$RSS(w) \equiv \sum_{i=1}^{n} \left\| \vec{x}_i - \sum_{j \neq i} w_{ij} \vec{x}_j \right\|^2$$

۳. مختصات Y را به صورتی مشخص می شود که بازسازی نقاط توسط وزن‌ها، حداقل گردد.

**آنالیز اجزای مستقل**[1]. آنالیز اجزای مستقل یک روش پردازش سیگنال آماری است کـه بـرای یافتن منبع‌های مستقل (تنها با در دست داشتن داده‌هایی که ترکیـب منـابع ناشـناخته هسـتند و بدون داشتن اطلاعات از مکانیزم ترکیب) طراحی شده است.

در این مدل داده‌های مشاهده‌شده به صورت X=AS نمایش داده می‌شود که در آن A مـاتریس ناشناخته‌ای است که نحوه‌ی ترکیب را نشان می‌دهـد و مـاتریس ترکیـب نامیـده مـی‌شـود و S منابعی هستند که به طور مستقیم از روی X قابل‌مشاهده نیستند.

مدل آنالیز اجزای مستقل نشان می‌دهد که داده‌های مشاهده‌شده X از ترکیب خطـی منـابع مکنون S حاصل شده و ماتریس ترکیب A چگونگی ترکیب خطی را نشان می‌دهد.

فرض اصلی این مدل این است که منابع مکنون از نظر آماری مستقل از یکدیگرند. بر پایـه ایـن فرض طی یک فرایند یادگیری بدون نظارت[2] آنالیز اجزای مستقل به کشف ماتریس تفکیـک W می‌رسد.

بعد از محاسبه ماتریس تفکیک W از آن برای انتقال داده‌هـای مشـاهده‌شـده X بـه منـابع مستقل Y استفاده می‌کند (WX=Y) و این منابع مسـتقل Y تخمینـی بـرای منـابع مکنـون S محسوب می‌شوند. به ردیف‌های ماتریس Y اجزای مستقل[3] گفته می‌شود و تا جایی کـه امکـان دارد باید نسبت به هم مستقل باشند.

هرچند از آنالیز اجزای مستقل در حل مسائل پـردازش سـیگنال در امـور پزشـکی، در زمینـه پردازش سیگنال صوتی، تشخیص چهره استفاده زیادی شده است، اما کمتر در پیش‌بینـی تقاضـا به کار رفته است.

بک و ویگند [۱۲] از آنالیز اجزای مستقل برای کشف متغیرهای اصلی در تعیین سود روزانـه ۲۸ سهام بزرگ ژاپن استفاده کرده‌اند، کیوی‌لوتو و اُجا [۱۳] ایـن آنـالیز را بـرای کشـف عوامـل

1. Independent Component Analysis
1. Unsupervised Learning
2. Independent Components

اصلی تعیین‌کننده گردش مالی ۴۰ مغازه در یک مجتمع تجاری به کار برده‌اند و اُجـا و همکـاران [۱۴] از آن برای پیش‌بینی نرخ تبادل ارز استفاده کرده‌اند.

**الگوریتم خوشه‌بندی کامینز**[1]. الگوریتم خوشه‌بندی کامینز، داده‌ها را بر مبنای ویژگی‌هایشان به K خوشه جداگانه تقسیم می‌کند. اشیائی که در یک خوشه قرار می‌گیرند، ویژگی‌های مشابه دارند. در زیر مراحل الگوریتم خوشه‌بندی کامینز آمده است:

۱. تعداد خوشه‌ها (K) مشخص می شود.
۲. مرکز ثقل برای خوشه‌ها تعیین می گردد.
۳. حلقه‌ای روی تمام اشیاء اجرا نموده و فاصله هر شیء را تا مرکز ثقـل همـه خوشـه‌هـا انـدازه گرفته شده و هر شیء به خوشه نزدیک‌ترین مرکز ثقل نسبت داده می شود.
۴. مراکز ثقل خوشه‌های جدید مجدداً محاسبه می گردد.
۵. مرحله ۳ مجدداً اجرا شده تا زمانی که مراکز ثقل دیگر تغییر نکنند.

در این الگوریتم فاصله بین دو شیء توسط تابع فاصله اقلیدسی (رابطه ۳) محاسبه می‌شود.

$$d(x,y) = \sqrt[2]{\sum_{i=1}^{m}(x_i - y_i)^2} \qquad \text{رابطه ۳}$$

تی و کاو [۵] مدل پیش‌بینی سری زمانی را با ترکیب SVM و خوشه‌بندی ارائه دادنـد. کـاو [۱۵] SVM و خوشه‌بندی را برای مسائل پیش‌بینی سـری زمـانی بـه کـار بـرده اسـت. لای و همکاران [۱۶] از الگوریتم K-means به همراه درخت تصمیم‌گیری فازی برای پیش‌بینی قیمـت سهام استفاده کرده‌اند. هوانگ و سای [۱۷] از مدل هیبریدی شامل مراحل خوشه‌بنـدی، SVR و انتخاب متغیر بر مبنای فیلتر برای کاهش زمان برازش مدل و بهبود دقت مدل استفاده کرده‌اند.

**ماشین بردار پشتیبان**[2] **و رگرسیون بردار پشتیبان**[3]. ماشین بردار پشتیبان یکی از روش‌های یادگیری با نظارتی[4] است که برای طبقه‌بندی و رگرسیون استفاده می‌شود. ماشین بردار پشتیبان در سال ۱۹۹۲ توسط وپنیک[5] معرفی شده و بر پایه تئوری یادگیری آماری بنا

1. K-Means
2. Support Vector Machines
3. Support Vector Regression
4. Supervised Learning
5. Vapn[ik]

نهاده شده است. ماشین بردار پشتیبان در انواع دسته‌بندی‌ها همچون تشخیص ارقام دستنویس، تشخیص ، شناسایی صورت، دسته‌بندی انواع صداها و مانند آن مورد استفاده قرار گرفته است که در مقایسه با تکنیک‌های دیگر از کارایی قابل توجهی برخوردار است. رویکرد SVM به این صورت است که در فاز آموزش، تلاش می‌شود مرز تصمیم‌گیری به گونه‌ای انتخاب شود که حداقل فاصله آن با هریک از دسته‌های موردنظر حداکثر شود. این نوع انتخاب باعث می‌شود که تصمیم‌گیری در عمل، شرایط نویزی را به خوبی تحمل کرده و پاسخ‌دهی خوبی داشته باشد. این نحوه انتخاب مرز بر اساس نقاطی به نام بردارهای پشتیبان انجام می‌شود. تفاوت اساسی این طبقه‌بندی‌کننده با سایر طبقه‌بندی‌کننده‌های آماری این است که برای پردازش و طبقه‌بندی داده‌های ابرطیفی، دیگر نیازی به کاهش تعداد باندها نمی باشد. مسئله رگرسیون غیرخطی می‌تواند مانند مسئله بهینه‌سازی بیان شود.

رابطه ۴

$$Minimize \frac{۱}{۲}w^{۲}+C\sum_{i=۱}^{۱}(\xi_i+\xi_i^*)$$

با لحاظ محدودیت‌های زیر

رابطه ۵

$$y_i-(w.\varphi(x_i)+b)\le\varepsilon+\xi_i$$
$$(w.\varphi(x_i)+b)-y_i\le\varepsilon+\xi_i^*$$
$$\xi_i,\xi_i^*\ge ۰, i=۱,۲,...,۱$$

بنابراین مسئله بهینه‌سازی در حالت غیرخطی به صورت زیر خواهد بود:

رابطه ۶

$$Minimize \frac{۱}{۲}\sum_{ij}^{۱}(a_i-a_i^*)(a_j-a_j^*)\langle\varphi(x_i).\varphi(x_j)\rangle$$
$$+\varepsilon\sum_{i=۱}^{۱}(a_i+a_i^*)-\sum_{i=۱}^{۱}y_i(a_i-a_i^*)$$

با لحاظ محدودیت‌های زیر

رابطه ۷

$$\sum_{i=1}^{l}(a_i - a_i^*) = 0$$
$$0 \le a_i \le C, i = 1,2,...,l$$
$$0 \le a_i^* \le C, i = 1,2,...,l$$

پارامترهای حاکم بر SVR غیرخطی، مقدار ثابت C ، شعاع لوله غیرحساس ε و پارامترهای هسته‌ای هستند. این پارامترها به طور متقابل به هم وابسته بوده و تغییر مقدار از یک پارامتر، به پارامترهای دیگر را تغییر می‌دهد. پارامتر C، شفافیت تابع تقریب را کنترل می‌کند. مقدار C بزرگ‌تر، خطای بیشتری را ایجاد نموده و یادگیری را پیچیده‌تر می‌کند و یک مقدار C کوچک‌تر به خطاهایی منجر می‌شود که قابل قبول بوده لیکن ممکن است یادگیری با تقویت ضعیف را ایجاد کند.

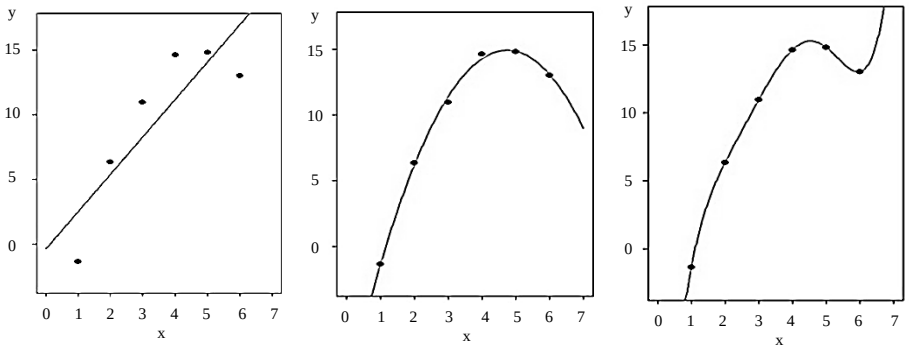
اگر تعداد داده‌ها زیاد شد، ممکن است مقادیر C کوچک‌تر که خطای کمتری را ایجاد می‌کند، ترجیح داده شود. پارامتر ε نیز بر پیچیدگی یا شفافیت تابع تقریب اثر می‌گذارد و حاکم بر تعداد بردارهای پشتیبان است. مقادیر کوچک‌تر ε ممکن است به بردارهای پشتیبان بیشتر و در نتیجه یک ماشین یادگیری پیچیده‌تر منجر شود و مقادیر بزرگ‌تر ε سبب می‌شود که ε-tube تعداد داده بیشتری را احاطه کند که در یادگیری در نظر گرفته نمی‌شود، بنابراین مقدار زیادی از اطلاعات مهم در داده‌ها از دست می‌رود.

رگرسیون بردار پشتیبان به طور موفقیت‌آمیز در مسائل پیش‌بینی سری زمانی و پیش بینی تقاضا استفاده شده است؛ برای نمونه در پیش‌بینی ارزش تولید ماشین‌آلات صنعتی، پیش‌بینی ضریب اطمینان موتور، پیش‌بینی سرعت باد و پیش‌بینی سری‌های زمانی مالی [۵، ۱۸، ۱۹، ۲۰، ۲۱، ۲۲، ۲۳ و ۲۴].

اگرچه استفاده از رگرسیون بردار پشتیبان در پیش‌بینی سری‌های زمانی مالی نتایج خوبی به همراه داشته است، تحقیقات کمی در بهره‌گیری از رگرسیون بردار پشتیبان برای پیش‌بینی تقاضا شده است.

**بیش برازش.** اگر به مدلی برای پیش‌بینی تقاضا، شامل X مقدار تقاضا و Y مبلغ تقاضا، به صورت شکل ۱ نیاز باشد، برای رسیدن به مدل مناسب پیش‌بینی تقاضا، از این ۶ نقطه رگرسیون گرفته می‌شود:

شکل ۱ نشان‌دهنده نتیجه سه رگرسیون متفاوت است؛ شکل سمت راست یک رگرسیون درجه ۵، شکل وسط رگرسیون درجه ۲ و شکل سمت چپ رگرسیون درجه ۱ را نشان می‌دهد.




*شکل۱. رگرسیون‌های مختلف بر اساس مقدار و مبلغ تقاضا*

همان طور که دیده می شود، شکل سمت راست (رگرسیون درجه ۵) از تمامی داده‌ها عبور می‌کند و بنابراین خطای نمونه این مدل صفر است. اگر چه این مدل، مدل مناسبی برای پیش‌بینی ۶ مقدار اندازه‌گرفته‌شده است، اما مدل مناسبی برای پیش‌بینی مقادیر خارج نیست؛ به سخن دیگر اگرچه مدل سمت راست خطای نمونه پایینی دارد، اما خطای عمومی آن بالا خواهد بود. مدل سمت چپ مدل بسیار ساده‌ای است که انتظار می‌رود خطای عمومی آن نیز بالا باشد. گرچه با اینکه خطای عمومی هر دو مدل سمت راست و چپ بالا است، اما خطای بالای آن‌ها به دلایل کاملاً متفاوتی بر می گردد.

در صورتی که رابطه اصلی میان x و y یک رابطه غیرخطی باشد، حتی اگر داده‌های بسیار زیادی در دست باشد، باز هم مدل سمت چپ (رگرسیون یک‌جمله‌ای) نخواهد توانست ساختار موجود در داده‌ها را کشف کند.

از طرف دیگر مدل های پیچیده سمت راست (رگرسیون پنج‌جمله‌ای)، در داده‌های کم، ساختارهایی را که به‌صورت تصادفی و بر حسب اتفاق در داده‌های ما وجود دارد، به اشتباه به عنوان ساختار عمومی رابطه x و y بیان خواهد کرد.

خطای بالای عمومی مدل سمت راست به دلیل استفاده کردن از مدلی بسیار پیچیده‌تر از حد مطلوب است و به اصطلاح «بیش‌برازش[1]» رخ داده است.

از سوی دیگر، مدل سمت چپ بسیار ساده است و به همین دلیل خطای عمومی بالایی دارد و به اصطلاح در مدل سمت چپ «کم‌برازش[2]» رخ داده است.

مدل وسط (رگرسیون درجه ۲) از میزان پیچیدگی مناسبی برخوردار است و به‌طور کلی میزان خطای عمومی به پیچیدگی مدل بستگی دارد.

1. Over fitting
2. Under fitting

با افزایش پیچیدگی مدل، خطا در نمونه (به‌طور پیوسته) کاهش یافته و مدل‌های پیچیده‌تر، خطا در نمونه کمتری دارند، اما این وضعیت برای خطای عمومی صادق نیست. همچنین با افزایش پیچیدگی تا مقدار خاصی، خطای عمومی کاهش می‌یابد، اما اگر پیچیدگی مدل از این مقدار معین بیشتر شود، خطای عمومی بالا خواهد رفت، زیرا با افزایش پیچیدگی به جای کشف روابط حقیقی میان متغیرها، اغتشاش موجود در داده‌ها پردازش می‌شود و مدل بیش‌برازش می‌شود.

همچنین مدل‌هایی که پیچیدگی کمتری از میزان بهینه دارند، مشکل کم‌برازش دارند. لازم به ذکر است که موضوع خطا در این زمینه از اهمیت زیادی برخوردار است. شناخت اجزای اصلی خطای کلی کمک زیادی به حل موضوع می نماید.

به‌طور کلی خطای کلی یک مدل به سه جزء زیر تقسیم می‌شود:

رابطه ۸ (خطای کلی) = ( نویز) + (بایاس) + (واریانس)

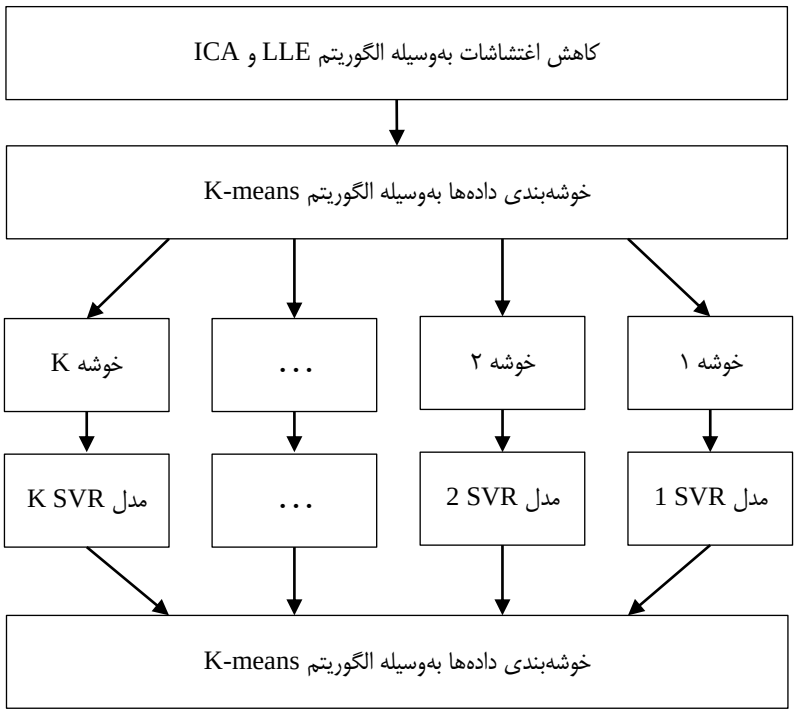
جزء نویز[1] در اصل جزء غیرقابل پیش‌بینی مسئله محسوب می‌شود و شامل مقدار اغتشاشات آماری[2] است که حتی در اطراف بهترین پیش‌بینی ها وجود دارد، جزء بایاس[3] باعث می‌شود خطای عمومی در مدل‌هایی که از کم‌برازشی رنج می‌برند، افزایش یابد و جزء واریانس[4] باعث می‌شود خطای عمومی در مدل‌هایی که از بیش‌برازشی رنج می‌برند، افزایش یابد.

**مدل مفهومی.** مدل پیشنهادی این مقاله یک مدل ترکیبی[5] و شامل سه مرحله کاهش بعد، خوشه‌بندی و پیش‌بینی است.

در مرحله اول تلاش شده با کاهش بعد داده‌ها اغتشاش‌های موجود در آن‌ها تا حد امکان گرفته شده و در مرحله دوم با دسته‌بندی و خوشه‌بندی کردن داده‌ها و قرار دادن داده‌های مشابه در گروه‌های یکسان، دقت پیش‌بینی بالا خواهد رفت و در مرحله‌ی سوم پیش‌بینی روی داده‌هایی انجام می‌شود که اغتشاش‌های آن‌ها تا حد امکان گرفته شده است و دسته‌بندی شده‌اند.

1. Noise
2. Statistical Fluctuation
3. Bias
4. Variance
5. Hybrid Model

مراحل اجرای کار در شکل ۲ مشاهده می شود. همان‌طور که در این شکل مشخص است، بعد از اجرای مراحل سه‌گانه فوق، به تعداد دسته‌های به‌دست آمده از مرحله خوشه‌بندی، مدل پیش‌بینی‌کننده وجود خواهد داشت.



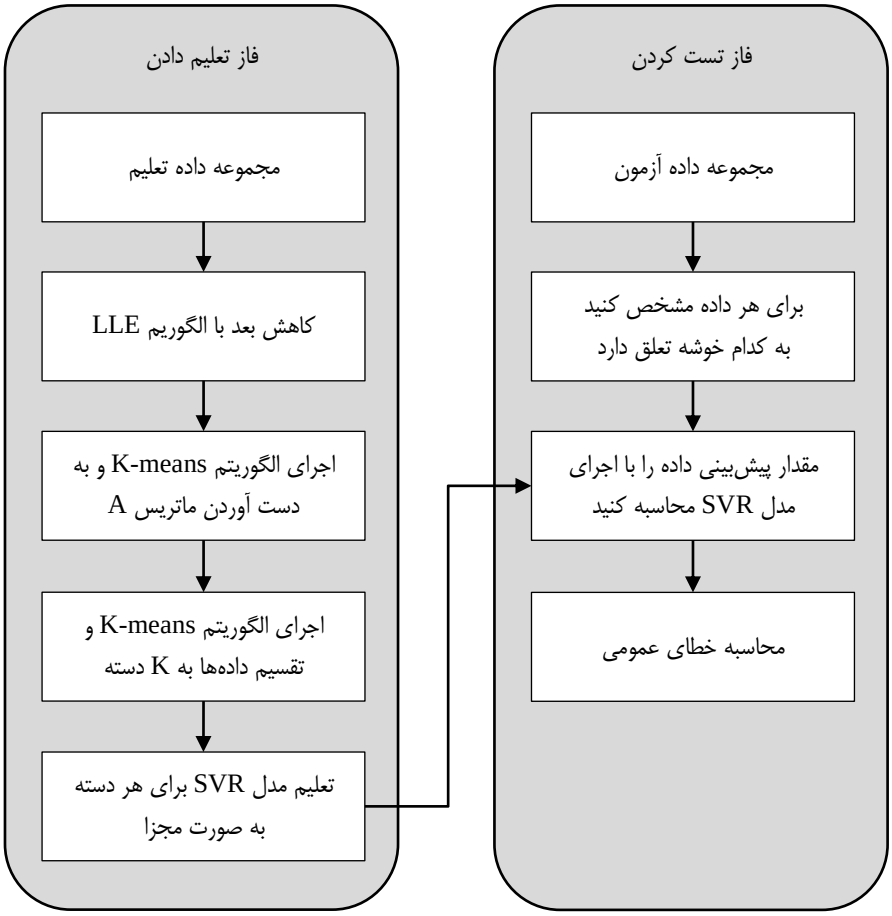

*شکل ۲. مراحل اجرای کاهش اغتشاش‌ها*

مراحل ذکرشده در بالا باید برای دستیابی به مدل‌های پیش‌بینی‌کننده انجام شوند، اما اگر بخواهیم بعد از به دست آوردن مدل‌های پیش‌بینی، میزان فروش داده‌های جدید را پیش‌بینی کنیم، باید گام‌های زیر برداشته شود:

۱. ابتدا باید مشخص شود که داده‌ای که خواهان پیش‌بینی آن هستیم به کدام‌یک از دسته‌های ۱ تا K تعلق دارد.

۲. بعد از مشخص کردن دسته موردنظر، از مدل مربوط به آن دسته برای انجام پیش‌بینی استفاده خواهد شد.

نحوه اجرای پیش‌بینی در فاز آزمایشی در شکل ۳ نشان داده شده است.




*شکل۳ . مراحل تعلیم دادن و آزمایش کردن داده‌ها برای پیش‌بینی*

## ۳. روش‌شناسی

**داده‌های به‌کارگرفته شده.** شرکت سهامی عام کارخانه چینی ایران (کاشی ایرانا) اولین و بزرگترین تولیدکننده کاشی دیواری به روش صنعتی در ایران است که در سال ۱۳۳۶ با سرمایه ۵۲۰ میلیون ریال تحت شماره ۵۹۰۴ به ثبت رسیده است. این شرکت اکنون با داشتن ۴ سالن تولید، انواع کاشی دیواری به ابعاد ۴۵ در۳۰، ۶۰ در ۳۰ و ۴۰ در ۲۵ و کاشی کف به ابعاد ۵۰ در ۵۰، ۴۰ در ۴۰ ، ۳۳ در ۶۰ و ۲۵ در ۵۰ را تولید می‌کند.

در تحقیق حاضر از اطلاعات واحد فروش ۵۰ مورد فروش، در طی ۳ سال و به تفکیک ماه استفاده شده است.

این اطلاعات را می‌توان به صورت یک ماتریس در نظر گرفت؛ به این صورت که هر سطر از این ماتریس، اطلاعات فروش یک مدل و هر ستون، اطلاعات فروش یک ماه را نشان می‌دهد. از آنجا که در این پژوهش ۵۰ مورد بررسی شده است، لذا ۵۰ سطر خواهیم داشت و از آنجا که از اطلاعات ۳۶ ماه استفاده شده است، ۳۶ ستون خواهیم داشت؛ بنابراین ماتریس به‌دست آمده، یک ماتریس ۳۶ در ۵۰ خواهد بود.

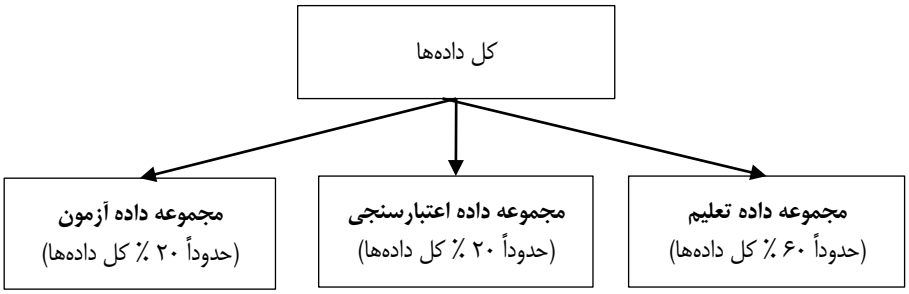
با توجه به اینکه این ماتریس ۱۸۰ درایه مرتبط با ۱۸۰ مقدار فروش دارد. برای پیش‌بینی فروش هر ماه از پنج متغیر زیر استفاده می شود:

۱. فروش یک ماه قبل؛ ۲. فروش دو ماه قبل؛ ۳. فروش سه ماه قبل؛ ۴. میانگین متحرک[1] سه‌ماهه و ۵. میانگین متحرک شش‌ماهه.

ماتریس دوم به این صورت ایجاد می‌شود که در ستون اول مقادیر فروش را درج نموده و در ستون‌های ۲ تا ۶، پنج متغیر بالا آورده می شود. از آنجا که ۱۸۰ داده فروش وجود داشته و متناظر هر داده پنج متغیر پیش‌بینی‌کننده موجود است، ماتریس به دست آمده، یک ماتریس ۶ در ۱۸۰ خواهد بود.

پس از حذف سطرهایی که دارای درایه‌ای خالی هستند، اندازه ماتریس حاصل شده ۶ در ۱۵۰ خواهد بود.

کل داده‌های موجود را مطابق شکل ۴ تقسیم نموده و تعریف هریک از مجموعه‌ها و دلایل تقسیم‌بندی مزبور در ادامه آورده شده است.




*شکل ۴. مجموعه داده‌ها*

1. Moving Average

مجموعه داده تعلیم:

مجموعه داده تعلیم داده‌هایی هستند که در برازش مدل به کار می‌روند. در این تحقیق حدوداً ۶۰ درصد داده‌های ابتدایی به عنوان مجموعه داده تعلیم انتخاب شده است.

مجموعه داده اعتبارسنجی:

باید در انتخاب مدل، مقادیری برای پارامترهای مدل تعیین شود؛ برای نمونه، در الگوریتم SVM همان‌طور که در قبل توضیح داده شد، باید مقادیر بهینه‌ای برای سه پارامتر C، $\Sigma$ و $\Upsilon$ انتخاب شود زیرا تغییر این پارامترها تأثیر زیادی بر نتایج کار و دقت پیش‌بینی می‌گذارد.

روشن است اگر تنها از مجموعه داده‌های تعلیم برای تعیین مقادیر بهینه C، $\Sigma$ و $\Upsilon$ استفاده شود، این خطر وجود دارد که مقادیر انتخاب‌شده تنها مقادیر بهینه برای همان مجموعه باشند؛ به سخن دیگر، با این کار خطا در نمونه پایین آمده، اما خطای عمومی کاهش نیافته است.

به این ترتیب، برای انتخاب پارامترهای بهینه، مدل نیازمند داده‌هایی است که در برازش مدل از آنها استفاده شده است. به این داده‌ها مجموعه داده اعتبارسنجی گفته می‌شود که در تحقیق حاضر حدوداً ۲۰ درصد از کل داده‌ها را به عنوان مجموعه داده اعتبارسنجی انتخاب شده است.

به این روش انتخاب پارامترها، اعتبارسنجی ضربدری[1] گفته می‌شود و گام‌های اجرای آن بعد از تفکیک داده‌ها به صورت زیر است:

۱. مجموعه‌هایی از مقادیر پارامترها را انتخاب می‌کنیم؛ برای نمونه، در مثال SVM ۱۰ مجموعه از مقادیر C، $\Sigma$ و$\gamma$ را به صورت زیر انتخاب می‌شود:

P1: $(C_1, \sum_1, \gamma_1)$
P2: $(C_2, \sum_2, \gamma_2)$
P3: $(C_3, \sum_3, \gamma_3)$
P10: $(C_{10}, \sum_{10}, \gamma_{10})$

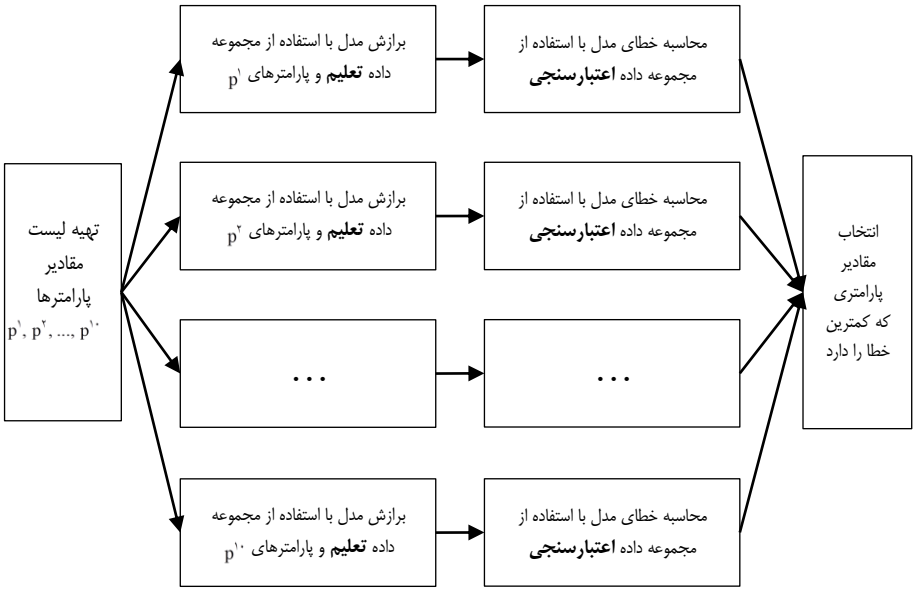
۲. با استفاده از مجموعه داده‌های تعلیم، مدل در هریک از این ۱۰ حالت برازش می‌شود، در نتیجه به ۱۰ مدل متفاوت دست یافته می شود.

۳. خطای این ۱۰ مدل را با مجموعه داده‌های اعتبارسنجی مورد سنجش قرار داده می شود.

۴. مجموعه مقادیر پارامتری را که کمترین خطا را داشته است، به عنوان پارامترهای بهینه مدل انتخاب می‌گردد.

در شکل ۵ الگوریتم اجرای گام‌های فوق مشاهده می شود:

1. Cross Validation




*شکل ۵. مراحل اجرایی*

در عمل ممکن است برای هریک از پارامترها ۱۰ مقدار متفاوت انتخاب شده و تمامی مجموعه‌های سه‌تایی از آن‌ها به عنوان P های شکل بالا در نظر گرفته شود؛ در این صورت در کل $۱۰^۳ =$ ۱۰۰۰ مجموعه خواهیم داشت و انتخاب پارامتر بهینه از میان این مجموعه‌ها انجام می‌گیرد.

**مجموعه داده آزمون.** در نهایت برای ارزیابی عملکرد مدل،به مجموعه داده‌های جدیدی نیاز می باشد. با توجه به اینکه مدل پیش‌بینی مطلوب باید عملکرد خوبی روی داده‌های جدید داشته باشد.بنابر این اگر از مجموعه داده‌های تعلیم یا اعتبارسنجی برای سنجش عملکرد مدل تحقیق استفاده شود، سنجش درست نبوده و تخمین ما از عملکرد مدل بهتر از عملکرد واقعی آن خواهد بود، زیرا در برازش مدل از دو مجموعه ذکرشده استفاده شده و به اصطلاح، مدل داده‌های این دو مجموعه را دیده است.

به این ترتیب، به مجموعه داده‌هایی نیاز داریم که در هیچ مرحله‌ای از آن استفاده نشده باشد. به این مجموعه، مجموعه داده آزمون گفته می‌شود و تنها یک بار در انتهای کار و برای

سنجش عملکرد واقعی مدل (سنجش خطای عمومی[1]) از آن استفاده می‌شود. در این مقاله حدوداً ۲۰ درصد داده‌های انتهایی را به‌عنوان مجموعه داده تعلیم انتخاب گردیده است.

## ٤. تجزیه و تحلیل داده‌ها

برای ارزیابی عملکرد مدل پیشنهادی این تحقیق، دقت پیش‌بینی مدل را با مدل‌های رگرسیون خطی و مدل SVR ساده مقایسه شده است.

لازم به ذکر است که قبل از اجرای رگرسیون خطی، برای از بین بردن همبستگی متغیرها از آنالیز اجزای اصلی استفاده گردید و پارامترهای بهینه مدل SVR ساده را از طریق اعتبارسنجی ضربدری به دست آمده است.

همچنین برای محاسبه خطای مدل‌ها از مجذور میانگین مربعات خطا[2] استفاده شده است. نحوه محاسبه این خطا را در رابطه ۹ مشاهده می شود:

$$RMSE = \sqrt[2]{\frac{\sum_{i=1}^{n}(T_i - P_i)^2}{n}}$$ رابطه ۹

که در آن T مقدار حقیقی، P مقدار پیش‌بینی‌شده و n تعداد کل نقاط می باشد.

عملکرد این سه مدل در جدول ۱ مشاهده می شود. نتایج نشان می‌دهد که مدل پیشنهادی این تحقیق کمترین خطای عمومی را از میان مدل‌های دیگر دارا بوده و بیانگر عملکرد بهتر این مدل می باشد.

| | نتایج | |
|---|---|---|
| | خطا در نمونه | خطای عمومی |
| رگرسیون خطی | ۰/۳۶۱۰ | ۰/۸۴۰۶ |
| SVR | ۰/۴۵۰۹ | ۰/۷۵۷۰ |
| مدل پیشنهادی | ۰/۵۷۳۷ | ۰/۶۳۹۹ |

*جدول ۱. مقایسه عملکرد مدل‌ها*

1. Generalization error
1. Root mean square error (RMSE)

## ۵. نتیجه‌گیری

در مقاله حاضر، پیش بینی فروش کاشی و سرامیک به کمک رگرسیون بردار پشتیبان انجام شده و نشان داده شده است که پیش بینی بوسیله رگرسیون مزبور در مسائل کاربردی از نتایج دقیق تری برخوردار می باشد. از آنجایی که وجود اغتشاش در داده ها نتایج پیش بینی را به مقدار زیادی تحت تأثیر قرار می دهد، لذا پیش پردازش داده ها از اهمیت زیادی برخوردار است. در این پژوهش با کاهش بعد داده ها و با پیشنهاد استفاده از آنالیز مولفه های مستقل و یادگیری منیفلد تا حد قابل ملاحظه ای اغتشاشات موجود در داده های مورد مطالعه کاهش یافته است.

از جمله روش های کاهش خطای عمومی در مقابل کاهش خطا در نمونه، استفاده از روش اعتبار سنجی ضربدری است که در مقاله حاضر از این روش جهت دستیابی به پارامترهای بهینه مدل استفاده شده است. همانگونه که در جدول شماره ۱ نشان داده شده است، مدل پیشنهادی این پژوهش برای پیش بینی در مقایسه با برخی از مدلهای بررسی شده، بدلیل دارا بودن خطای عمومی کمتر از دقت بالاتری برخوردار می باشد. بدین ترتیب برای پیش بینی تقاضا در صنعت کاشی و سرامیک توصیه می شود از مدل پیشنهادی این پژوهش استفاده شود.

**پیشنهاد برای تحقیقات آتی.** برای تحقیقات آتی موارد زیر پیشنهادی می شود:

ـ بررسی مدل با اضافه کردن متغیرهای دیگر به متغیرهای پیش‌بینی مدل تحقیق مثل اطلاعات رقبا، اطلاعات محیط صنعت؛

ـ بهره‌گیری از مفاهیم آمار بیزی در طراحی مدل‌های پیش‌بینی؛

ـ طراحی مدل‌های جدید با مجموعه‌ای از داده‌های بیشتر برای زمان‌های دورتر و برای سایر صنایع.

## منابع

1. Delaney, R., & R. Wilson. (2003). *14th Annual State of Logistics Report.*

2. Simchi-Levi, D. ( 2004). *Managing The Supply Chain.* New York :McGraw-Hill

3. Henkoff, R. (1994). Delivering the Goods. *Fortune,* 64–78

4. Tay, F.E.H., & Cao, L.J. (2003). Support vector machine with adaptive parameters in financial time series forecasting. *IEEE Transactions on Neural Networks 14,* 1506–1518.

5. Tay, F.E.H., & Cao, L.J. (2001). Improved financial time series forecasting by combining support vector machines with self-organizing feature map. *Intelligent Data Analysis 5,* 339–354.

6. Chang, P.C., Wang, Y.W., & Tsai, C.Y. (2005). Evolving neural network for printed circuit board sales. *Expert Systems with Applications 29 (1),* 83–92.

7. Vapnik, V.N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks 10,* 988–999.

8. Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory.* New York :Springer.

9. Chang, P.C., Wang, Y.W., & Liu, C.H. (2006). Combining SOM and GA-CBR for flow time prediction in semiconductor manufacturing factory. *Lecture Notes in Compu- ter Science 4259,* 767–775.

10. Wang, Y.W., Liu, C.H., & Fan, C.Y. (2009). The hybrid model development of clustering and back propagation network in printed circuit board sales forecasting. *Opportunities and Challenges for Next-Generation Applied Intelligence 214,* 213–218.

11. Chi, J.L., & Yen, W.W. (2010). Combining independent component analysis and growing hierarchical self-organizing maps with support vector regression in product demand forecasting. *Int. J. Production Economics 128,* 603-613.

12. Back, A., & Weigend, (1997). Discovering structure in finance using independent component analysis. *Proceedings of the Fifth International Conference on Neural Networks in Capital Market, Kluwer Academic,* pp. 15–17.

13. Kiviluoto, K., & Oja, E. (1998). Independent component analysis for parallel financial time series. *Proceedings of the Fifth International Conference on Neural Information, Tokyo, Japan,* pp. 895–898.

14. Oja, E., Kiviluoto, K., & Malaroiu, S. (2000). Independent component analysis for financial time series. *Proceeding of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium, Lake Louise, Canada, pp*. 111–116.

15. Cao, L.J. (2003). Support vector machines experts for time series forecasting. *Neurocomputing 51,* 321–339.

16. Lai, R.K., Fan, C.Y., Huang, W.H., & Chang, P.C. (2009). Evolving and clustering fuzzy decision tree for financial time series data forecasting. *Expert Systems with Applications 36 (2 PART 2)*, 3761–3773.

17. Hung, C., & Tsai, C.F. (2008). Market segmentation based on hierarchical self- organizing map for markets of multimedia on demand. *Expert Systems with Applications 34*, 780–787.

18. Mohandes, M.A., Halawani, T.O., Rehmam, S., & Hussain, A.A (2004). Support vector machines for wind speed prediction. *Renewable Energy 29*, 939–947.

19. Thissen, U., Brakel, R.V., De Weijer, A.P., Melssen, W.J., & Buydens, L.M.C. (2003). Using support vector machines for time series prediction. *Chemometrics and Intelligent Laboratory Systems 69*, 35–49.

20. Lu, C.-J., Lee, T.-S., & Chiu, C.-C. (2009). Financial time series forecasting using independent component analysis and support vector regression. *Decision Support Systems 47 (2)*, 115–125.

21. Pai, P.F., Yang, S.L., & Chang, P.T. (2009). Forecasting output of integrated circuit industry by support vector regression models with marriage honey-bees optimization algorithms. *Expert Systems with Applications 36 (7)*, 10746–10751.

22. Yang, Y., Fuli, R., Huiyou, C., & Zhijiao, X. (2007). SVR mathematical model and methods for sale prediction. *Journal of Systems Engineering and Electronics 18 (4)*, 769–773.

23. He, W., Wang, Z., & Jiang, H. (2008). Model optimizing and feature selecting for support vector regression in time series forecasting. *Neurocomputing 72 (1–3)*, 600–611.

24. Fang, S.F., Wang, M.P., Qi, W.H., & Zheng, F. (2008). Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials. *Computational Materials Science 44 (2)*, 647–655.