مهندسی کامپیوتر گرایش هوش مصنوعی

پایان‌نامه کارشناسی ارشد

مرکز آموزش الکترونیکی دانشگاه شاهرود

توسعه روشی جهت انتخاب نمادهای مؤثر در پیش‌بینی بازار بورس ایران

جواد خدامی

استاد راهنما:

دکتر علی سلیمانی ایوری

استاد مشاور:

مهندس فرزانه اکبر زاده

شهریور ۱۳۹۴

دانشگاه شاهرود
مدیریت تحصیلات تکمیلی
فرم شماره (۶)

باسمه تعالی

شماره: ۱۵۳۸/م/۱۰الف
تاریخ: ۹۶/۷/۲۸
ویرایش:

فرم صورت جلسه دفاع از پایان نامه تحصیلی دوره کارشناسی ارشد

با تأییدات خداوند متعال و با استعانت از حضرت ولی عصر (عج) نتیجه ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای حامد خدائی به شماره دانشجویی ۹۱۲۶۶۳۴ رشته نرم افزار کامپیوتر گرایش هوش مصنوعی تحت عنوان ترسیم روش جمع انتخاب نمادهای موثر در بازار بورس ایران که در تاریخ ۹۶/۶/۲۵ با حضور هیأت محترم داوران در دانشگاه شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

قبول (با درجه : بسیار خوب امتیاز ۱۸) ☑ دفاع مجدد ☐ مردود ☐

۱ـ عالی (۲۰ ـ ۱۹)
۲ـ بسیار خوب (۱۸/۹۹ ـ ۱۸)
۳ـ خوب (۱۷/۹۹ ـ ۱۶)
۴ـ قابل قبول (۱۵/۹۹ ـ ۱۴)
۵- نمره کمتر از ۱۴ غیر قابل قبول

اعضای داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱ـ استادراهنما	جناب دکتر علی سلیمانی ایدرگی	دانشیار	
۲ـ استاد مشاور	خانم مهندس فرزانه اکبرزاده		
۳ـ نماینده شورای تحصیلات تکمیلی	محسن فرهادی	ری	
۴ـ استاد ممتحن	جناب دکتر علی اکبر یزدیان	استاد	
۵ ـ استاد ممتحن	جناب دکتر مرتضی زاهدی		

رئیس دانشکده :

امضاء

تعهدنامه

اینجانب جواد خدامی دانشجوی دوره‌ی کارشناسی ارشد رشته‌ی کامپیوتر گرایش هوش مصنوعی دانشکده‌ی مجازی دانشگاه شاهرود نویسنده‌ی پایان‌نامه‌ی توسعه روشی جهت انتخاب نمادهای مؤثر در پیش‌بینی بازار بورس ایران تحت راهنمایی دکتر علی سلیمانی ایوری متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام‌شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققان دیگر به مرجع مورداستفاده استناد شده است.
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه شاهرود می‌باشد و مقالات مستخرج با نام «دانشگاه شاهرود» و یا « Shahrood University» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده‌اند در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در کلیه‌ی مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آن‌ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه‌ی مراحل انجام این پایان‌نامه، در مواردی که به حوزه‌ی اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه‌ی حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه شاهرود می‌باشد. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.

تقدیم به

پدر و مادرم،

اسطوره فداکاری و گذشت و مهربانی

تشکر و قدردانی:

آماده‌سازی یک رساله کارشناسی ارشد در زندگی حرفه‌ای و هم شخصی پر از تجربه‌های منحصربه‌فردی است که بدون یاری و کمک دیگران قابل انجام نیست و به سرانجام نمی‌رسید.

بسیار خرسندم که بدین‌وسیله مراتب قدردانی و احترام از استاد راهنمای بزرگوارم جناب آقای دکتر سلیمانی به خاطر راهنمایی‌ها و هدایت و دلگرمی در طول انجام این رساله سپاسگزاری نمایم، او که به من فرصت داد تا خود به دنبال مسیرم بگردم. همچنین خانم اکبر زاده و تمامی اساتید ارجمند به دلیل توجه به کار بنده و زمانی که صرف پایان‌نامه من نموده‌اند، بین‌هایت سپاسگزارم، شاید که مقبول افتد.

چکیده

تجزیه‌وتحلیل بازار مالی همیشه توجه زیادی از سرمایه‌گذاران و پژوهشگران را به خود جلب کرده است. قیمت بازار سهام بسیار پیچیده بوده و توسط عوامل مختلف تحت تأثیر قرار می‌گیرد. بنابراین پیدا کردن عوامل مهم و مؤثر در بازار سهام بسیار مهم است و از سوی دیگر پیش‌بینی آن بسیار مشکل می‌باشد، در مقالات اخیر بیشتر تلاش محققین در جهت پیش‌بینی بازار با کمک یکسری ویژگی‌ها معمولاً باهدف پیش‌بینی قیمت بازار بوده است، که نتایج حاصل از آن در بازار کاربرد چندانی نداشته است، یکی از روش‌های مهم جهت پیش‌بینی روند بازار استفاده از نشانگرها می‌باشد که تعداد این نشانگرها زیاد است و اینکه در آن بازه زمانی کدام‌یک مؤثر است اهمیت زیادی دارد. در این پژوهش ابتدا با توجه به نمودار قیمت سهم در یک بازه زمانی نشانگرهای مختلف محاسبه می‌گردند. به‌منظور دسته‌بندی روند قیمت به دودسته صعودی و نزولی از ماشین بردار پشتیبان استفاده گردید و برای شناسایی عوامل مؤثر از روش انتخاب ویژگی Wrapper رو به عقب استفاده گردید که نشانگرهای غیر مؤثر در صحت را حذف نموده و درنهایت نشانگرهای مؤثر باقی می‌مانند. با اجرای این الگوریتم بر روی سهام سیمان شاهرود و نفت پارس توانستیم موثرترین نشانگرها را با توجه به سری زمانی هر سهام بدست آوریم.

کلمات کلیدی: ماشین بردار پشتیبان، پیش‌بینی روند بورس، انتخاب ویژگی، نشانگر

فهرست مطالب

فهرست شکل‌ها

فهرست جدول‌ها

فصل اول

مقدمه

۱-۱ مقدمه

امروزه پیش‌بینی بورس[1] به علت متغیر بودن بازمان و وجود نویز در داده‌ها و نیز وجود ویژگی‌های زیاد تأثیرگذار در آن، یک چالش محسوب می‌شود. این پیش‌بینی شامل پیش‌بینی قیمت و پیش‌بینی روند بورس است که در اینجا به دنبال پیش‌بینی روند فردای بازار هستیم و برای این منظور از یک سری ویژگی‌های استخراج‌شده از قیمت استفاده می‌کنیم. منظور از روند، صعودی و یا نزولی بودن قیمت بازار است، یعنی قیمت بسته شدن فردای بازار در مقایسه با امروز اگر رشد داشته آن را دسته‌ی صعودی می‌نامیم و اگر نزول داشته آن را در دسته‌ی نزولی قرار می‌دهیم. در اینجا به دنبال این هستیم که بتوانیم با استفاده از دستگاه‌های هوشمند مقادیر تأثیرگذار را پیدا نموده و با استفاده از روش‌های سری زمانی[2] و دسته‌بندی[3] مسئله را حل نماییم.

پیش‌بینی روند بورس، به‌طور طبیعی برای یک تحلیلگر بازار سرمایه کار دشواری محسوب می‌شود، زیرا داده‌های بورس نویزی و متغیر بازمان و نیز متغیرهای زیادی در آن تأثیر دارند که کار پیش‌بینی را دشوار می‌کنند. به‌طورکلی محققین به پیش‌بینی روند بورس، قیمت و شاخص‌ها علاقه‌مند هستند، تعداد زیادی روش برای این منظور معرفی‌شده است. این روش‌ها شامل شبکه عصبی مصنوعی[4]، رگرسیون خطی[5] و چندخطی[6]، الگوریتم ژنتیک[7]، ماشین بردار پشتیبان[8] است.

مقاله‌های موجود درزمینه‌ی پیش‌بینی بورس را جمع‌بندی کردند و به این نتیجه رسیدند که روش‌های زیادی برای انتخاب ویژگی وجود دارد و ویژگی‌های متفاوتی نیز در مقاله‌های مختلف مورداستفاده

---

[1] Stock
[2] Time Series
[3] Classification
[4] Artificial neural network
[5] Linear regression
[6] Multi linear regression
[7] Genetic algorithm
[8] Support vector machine

قرارگرفته است. که معمولاً ورودی‌ها را عوامل نشانگر بازار یا تحلیل اخبار تشکیل می‌دهند. عمل انتخاب ویژگی به این خاطر انجام می‌شود تا بتوان به یک سری از ویژگی‌ها دست‌یافت که بتوانند به‌صورت بهتری داده‌ها را از هم جدا نمایند و صحت کار پیش‌بینی را بالا ببرند [۳۴]

داده‌های نشانگر بازار[1] با استفاده از یک سری فرمول از قیمت استخراج می‌شوند و نشان‌گر حالت حال بازار با توجه به گذشته‌ی قیمت هستند به‌طور مثال می‌توان میانگین متحرک را نام برد بدین شکل که میانگینی از قیمت بسته شدن گذشته‌ی بازار محاسبه می‌شود و یک دید در مورد قیمت آینده‌ی بازار می‌دهد، مثلاً بدین شکل که درصورتی‌که قیمت بالای میانگین باشد روند را صعودی نامیده و درصورتی‌که قیمت زیر میانگین باشد روند را نزولی در نظر می‌گیرند.

داده‌های اخبار به دودسته تقسیم می‌شوند دسته اول اخباری هستند که زمان اعلام آن‌ها مشخص است مثل نرخ بهره آمریکا و نرخ بیکاری و مثال‌هایی از این قبیل که روی قیمت ارز و سهم‌های گوناگون تأثیر می‌گذارند، دسته‌ی دوم اخبار لحظه‌ای هستند مانند ترور، زلزله و سخنرانی فرد خاص، که به‌صورت لحظه‌ای یا چندروزه روی بازارهای دنیا تأثیر می‌گذارند. پس ورودی را دو مورد تشکیل می‌دهند یا اخبار و یا ویژگی‌های استخراج‌شده از قیمت، در اینجا می‌خواهیم با استفاده از داده‌های نشانگر بازار به پیش‌بینی روند بازار بپردازیم.

مرور کارهای گذشته

در این فصل به پژوهش‌هایی که درزمینه‌ی پیش‌بینی سهام ارائه‌شده‌اند را بررسی کرده و از دو دیدگاه استخراج ویژگی و انتخاب ویژگی به‌عنوان ورودی به آن نگاه می‌کنیم و روش‌های ارائه‌شده را با یکدیگر مقایسه می‌نماییم.

[1] Indicator

۱-۲ پیش‌بینی روند سری زمانی بورس با استفاده از درخت تصمیم فازی

در این پژوهش[۱] از مجموعه داده‌ی سایت yahoo.com.tw استفاده کرده، و هدف آن پیدا کردن جهت روند با استفاده از تقسیم بازار به سه دسته‌ی Buy,Sell,Hold است. وCBFDT[1] نام‌گذاری شده است که با فرمول زیر قابل‌محاسبه است

$$\frac{X_i - X_{i-1}}{X_i} \qquad (۱-۱)$$

که در آن X_i نشان‌دهنده‌ی قیمت بسته شدن بازار است، نتیجه‌ی فرمول فوق اگر بزرگ‌تر از ۰,۵ باشد بازار صعودی در نظر گرفته‌شده و اگر کوچک‌تر از ۰,۵- باشد بازار نزولی، و اگر بین ۰,۵ و ۰,۵- باشد بازار Hold در نظر گرفته می‌شود. این سه مقدار دسته‌ها را تشکیل می‌دهند.

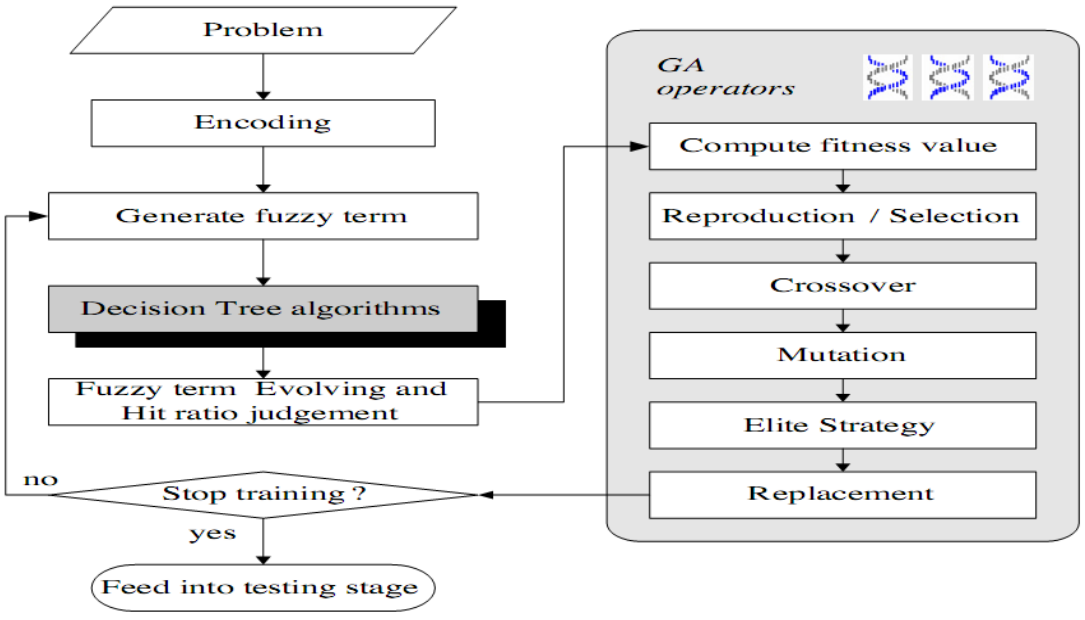
۱-۱- ویژگی‌هایی به اسم نشانگر بازار وجود دارند که با استفاده از فرمول‌هایی از قیمت، ویژگی‌هایی را برای نشان دادن جهت بازار استخراج می‌کنند. با کمک مرجع[۹] ویژگی‌هایی همچون RSI,KD,MACD,MA و Bias استخراج‌شده است. سپس با استفاده از روش Step-Wise regression analysis یک سری ویژگی که اثرگذاری بیشتری داشته‌اند را استخراج نموده‌اند که در جدول ۱-۱، ویژگی های استخراج شده به همراه ویژگی های انتخاب شده به همراه ویژگی های انتخاب شده برای هر سهام آمده است .

[1]Case based fuzzy decision tree

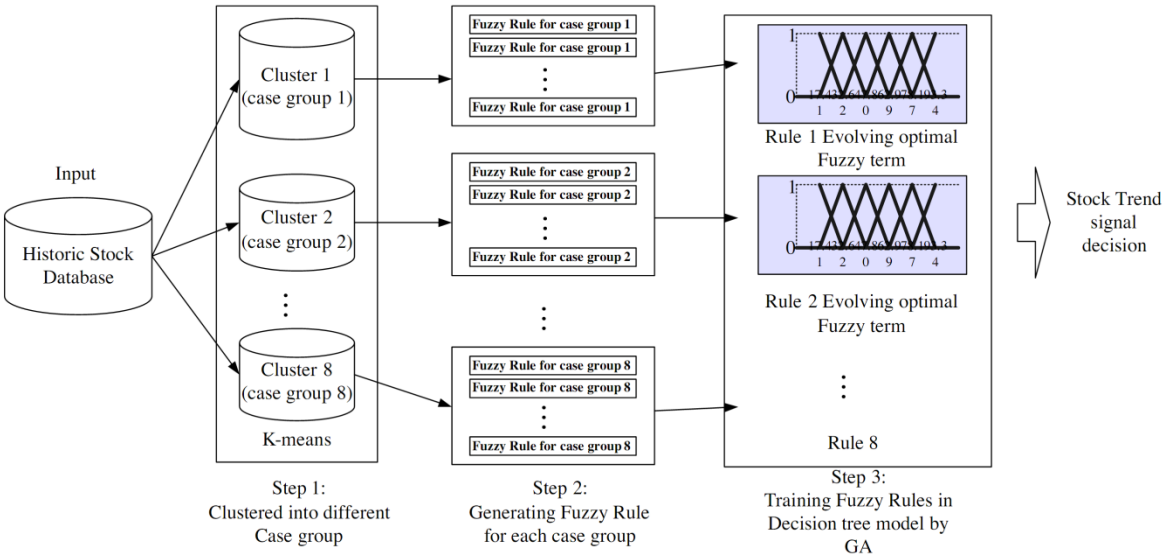
جدول۱-۱، ویژگی‌های استخراج‌شده به همراه ویژگی‌های انتخاب‌شده برای هر سهام

Input factors		
Total number of technical indices	5MA, 5BIAS, 6MA, 10MA, 9K, 9D, 6RSI, 9MACD, 12W%R, 20MA, 12RSI, K_D	
Total number of difference of technical indices	5MA difference, 6MA difference, 10MA difference, 5BIAS difference, 10BIAS difference, 6RSI difference, 9K difference, 9D difference, 12RSI difference, 12W%R difference	
Stock names	**Input factors**	
Google	Technical indices Difference of technical indices	12W%R 10BIAS difference 6RSI difference 12RSI difference
Yahoo	Technical indices Difference of technical indices	12RSI 12 W%R difference 6RSI difference 12RSI difference
IBM	Technical indices Difference of Technical indices	12W%R 10BIAS difference 6RSI difference 12 W%R difference
GM	Technical indices Difference of technical indices	12W%R 10BIAS difference 12RSI difference 12W%R difference
Microsoft	Technical indices Difference of technical indices	6RSI 10BIAS difference 12RSI difference 12 W%R difference

روش معرفی‌شده درخت تصمیم فازی است که توسط الگوریتم ژنتیک بهینه‌شده است که در شکل ۱-۱ آمده است.



شکل ۳-۱، ساختار آموزش درخت تصمیم فازی



شکل ۱-۲، ساختار پیشنهادی درخت تصمیم فازی

در این ساختار ابتدا داده‌ها به یک سری خوشه تقسیم‌شده و بعد قواعد فازی با استفاده از درخت تصمیم ساخته‌شده و درنهایت با استفاده از الگوریتم ژنتیک قواعد را بهبود می‌بخشد.

در جدول۱-۲، مقایسه‌ی روش پیشنهادی با روش‌های معرفی‌شده آمده که روش‌های متفاوت باهم مورد مقایسه قرارگرفته‌اند.

جدول۱-۲، مقایسه‌ی روش پیشنهادی با روش‌های معرفی‌شده درگذشته

Mining model	Hit ratios (%)
RW	51.06
ARIMA	56.13
BPNN	69.78
KBFDT (Lee & Park, 1989)	70.26
SVM (Kim, 2003)	78.65
ELSSVM (Lean et al., 2009)	82.66
GASVM (Yu et al., 2005)	84.57
CBRFDT	91.95

معیار مقایسه Hit Ratios در مقاله‌های زیادی موردمحاسبه قرارگرفته به همین دلیل این مقاله نیز از همین معیار استفاده کرده است. همان‌طور که از نتایج مشخص است روش پیشنهادی CBRFDT توانسته به صحت بهتری نسبت به روش‌های دیگر دست پیدا کند.

در این مقاله داده‌های دسته‌ها نامتقارن[1] بوده زیرا روند را به سه دسته تقسیم کرده بدین شکل که صعودی[2] نزولی[3] و نگه‌داشتن[4]، ۹۵ درصد داده‌ها شامل نگه‌داشتن و ۵ درصد فقط شامل صعودی و نزولی است (به‌طور مثال ۱۰۰۰۰۲۰۰۰۰۰۰۰۰۰۱۰۰۰۲ که در آن ۱ به معنای صعودی و ۰ نگه‌داشتن و ۲ نزولی می‌باشد) به همین دلیل یک دسته ۹۵ درصد داده‌ها را تشکیل داده و در این مقاله نیز فقط اگر کل داده‌ها را در دسته‌ی نگه‌داشتن قرار دهد به صحتی بالاتری نسبت به مقاله‌های قبل دست پیدا می‌کرد. و به همین دلیل باید نتایج هر دسته را به‌صورت جدا بیان می‌نمود تا مشخص می‌شد که چه تعداد از داده‌های هر دسته به‌درستی مورد پیش‌بینی قرارگرفته است.

۱-۳ پیش‌بینی سری زمانی سهام با استفاده از ساختار فازی

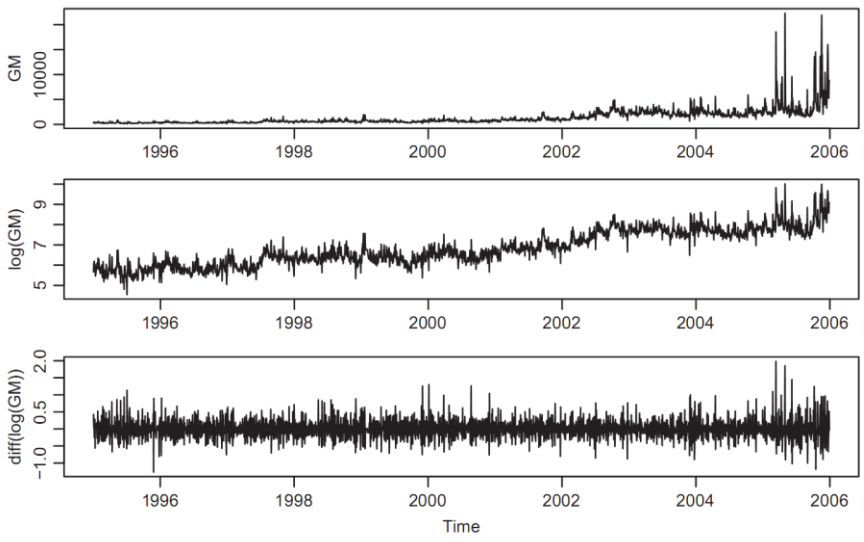
در مرجع [۱۰] روش‌های سنتی همچون AR[5] و TAR[6] و STAR[7] را با روش TSK[8] فازی مورد مقایسه قرار می‌دهد که ساختار بروزی به‌حساب نمی‌آید. مجموعه داده‌ی مورداستفاده بورس داو جونز می‌باشد که در جدول ۱-۲ ، سهام استفاده شده از بورس داو جونز آمده است.

[1] Imbalance
[2] Buy
[3] Sell
[4] Hold
[5] Autoregressive
[6] Threshold autoregressive
[7] Smooth transition autoregressive
[8] Takagi-Sugeno-Kang

جدول۱-۳، سهام استفاده‌شده از بورس داو جونز

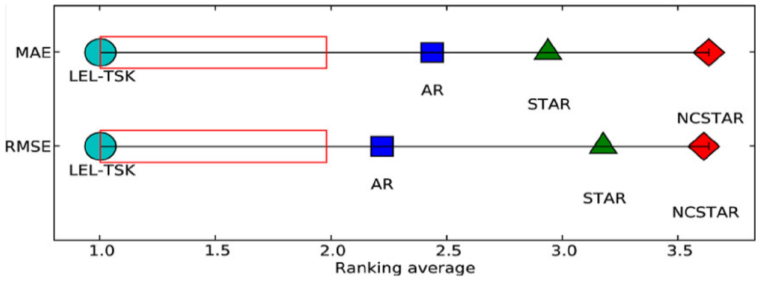
Code	Full name	Industry
aa	Alcoa	Aluminium
aig	American international group	Insurances
axp	American express	Consumer finance
ba	Boeing	Aerospace & defense
cat	Caterpillar	Construction & Mining equipment
dd	DuPont	Chemical industry
dis	Walt Disney	Broadcasting & Entertainment
ge	General electric	Conglomerate
gm	General motors	Automotive
hd	The home depot	Home improvement retailer
hom	Home solutions of America	Construction
ibm	IBM	Computers & technology
jnj	Johnson & Johnson	Pharmaceuticals
jpm	JPMorgan chase	Banking
ko	Coca-Cola	Beverages
mo	Altria group	Tobacco
mcd	McDonald's	Fast food
mmm	3 M	Conglomerate
mrk	Merck	Pharmaceuticals
pfe	Pfizer	Pharmaceuticals
pg	Procter & Gamble	Consumer goods
utx	United technologies corporation	Conglomerate
wmt	Wal-Mart	Retail

به‌منظور پیش‌بینی قیمت سهام، با استفاده از طول پنجره‌ی ۴، یعنی ۴ روز گذشته‌ی بازار به پیش‌بینی قیمت فردا می‌پردازد. برای این منظور ابتدا در داده‌ها تغییراتی اعمال می‌کند که در شکل۱-۳، ابتدا از قیمت سهام ژنرال موتور لگاریتم گرفته و سپس تفاوت دو روز را محاسبه نموده تا داده‌های بزرگ کوچک‌شده، سپس از این داده‌ها تفاضل گرفته بدین شکل که قیمت دو روز متوالی را از هم دیگر کم کرده و به مقدار جدیدی رسیده که میزان تفاوت دو روز می‌باشد، به‌طور مثال قیمت دیروز را از روز قبلش کم کرده است.



شکل۱-۳، ابتدا از قیمت سهام ژنرال موتور لگاریتم گرفته و سپس تفاوت دو روز را محاسبه نموده

به‌منظور آزمون روش و مقایسه از معیارهای خطای RMSE,MAE استفاده کرده و نتایج داده‌ی آزمون در شکل۱-۴، مقایسه‌ی نتایج روش فازی با روش‌های سنتی آمده است.



شکل۱-۴، مقایسه‌ی نتایج روش فازی با روش‌های سنتی

در جدول۱-۴، مقایسه‌ی روش فازی با روش‌های سنتی روی‌داده‌ی آموزش آمده است که میزان خطای روش فازی TSK به نسبت بقیه‌ی روش‌ها بالاتر بوده و نتایج خوبی به همراه نداشته است.

جدول۱-۴، مقایسه‌ی روش فازی با روش‌های سنتی روی‌داده‌ی آموزش

Model	Avg. ranking for RMSE	Avg. ranking for MAE
AR	4.00	3.98
STAR	2.22	2.17
NCSTAR	1.57	1.52
LEL-TSK	2.22	2.33

Nemenyi critical distance (p = 0.05): 0.978.

درحالی‌که در جدول۱-۵، مقایسه‌ی روش فازی با روش‌های سنتی روی‌داده‌ی آزمایش نشان‌دهنده‌ی این است که روش فازی روی‌داده‌های آزمایش توانسته به کمترین خطا دست پیدا کند.

جدول۱-۵، مقایسه‌ی روش فازی با روش‌های سنتی روی‌داده‌ی آزمایش

Model	Avg. ranking for RMSE	Avg. ranking for MAE
AR	2.22	2.43
STAR	3.17	2.93
NCSTAR	3.61	3.63
LEL-TSK	1.00	1.00

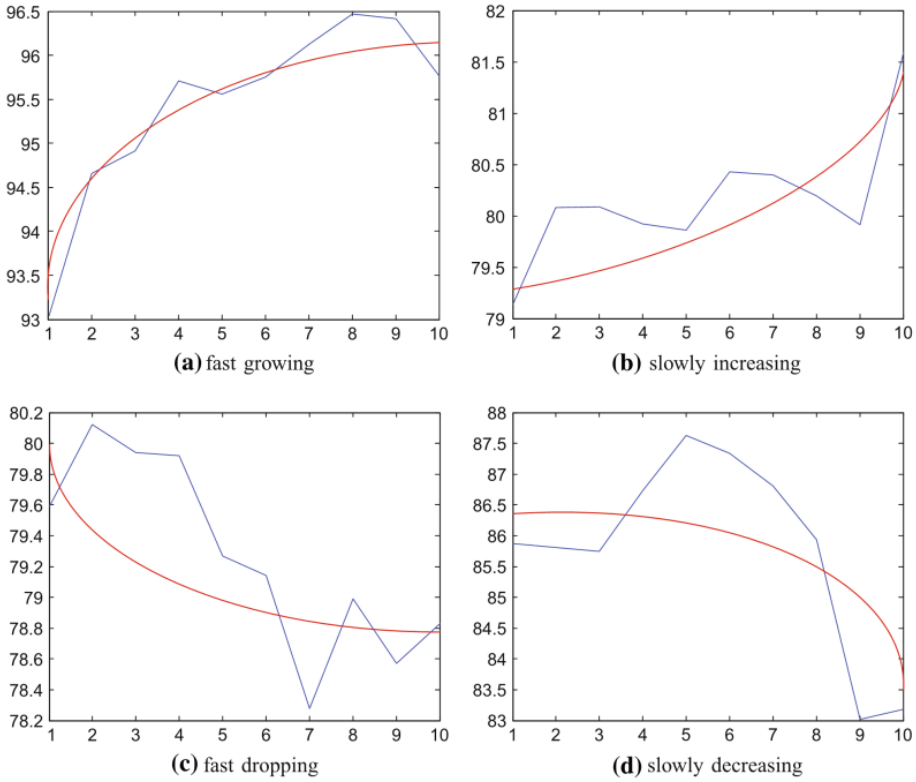
۱-۳-۱ نقدی بر پیش‌بینی سری زمانی سهام با استفاده از ساختار فازی

در این پژوهش ورودی فقط قیمت بوده و به همین دلیل توانایی پیش‌بینی قیمت پایین می‌آید. می‌توان از نشانگر بازارها نیز به‌عنوان ورودی نیز استفاده کرد و نیز ساختارهای مورد مقایسه‌ی استفاده‌شده قدیمی بوده و روش‌های بهتری نسبت به آن‌ها معرفی‌شده است.

۱-۴ پیش‌بینی سری زمانی فارکس، با استفاده از همبستگی ماشین بردار برازشی

در مرجع [۱۳] به بررسی تأثیر معیار همبستگی می‌پردازد. بدین شکل که با استفاده از رسم یک‌شکل گرافیکی و محاسبه‌ی مقدار همبستگی شکل باقیمت یک ویژگی استخراج کرده و به همراه قیمت به‌عنوان

ورودی به SVR[1] داده تا بتواند قیمت را پیش‌بینی کند.



شکل۱-۵، بازار به چهار حالت تقسیم‌شده

در شکل۱-۵، بازار به چهار حالت که این ۴ حالت نشان‌دهنده‌ی قدرت روند بازار می‌باشند، رشد سریع قیمت، رشد کند قیمت، نزول سریع قیمت، نزول کند قیمت.

حال نوبت به محاسبه‌ی همبستگی بین قیمت و این خطوط می‌رسد روش این پژوهش روش بهبود داده‌شده‌ی Pearson می‌باشد. و درنهایت ورودی در نظر گرفته‌شده قیمت به همراه مقدار همبستگی محاسبه‌شده می‌باشد.

روش مورداستفاده در این مقاله CSVR[2] می‌باشد، که ورودی آن را قیمت به همراه همبستگی قیمت و

[1] Support vector regression
[2] Correlation-aided support vector regression

چهار حالت تعریف‌شده تشکیل می‌دهند.

جدول۱-۶، مقایسه‌ی تأثیر ورود همبستگی به‌عنوان ورودی

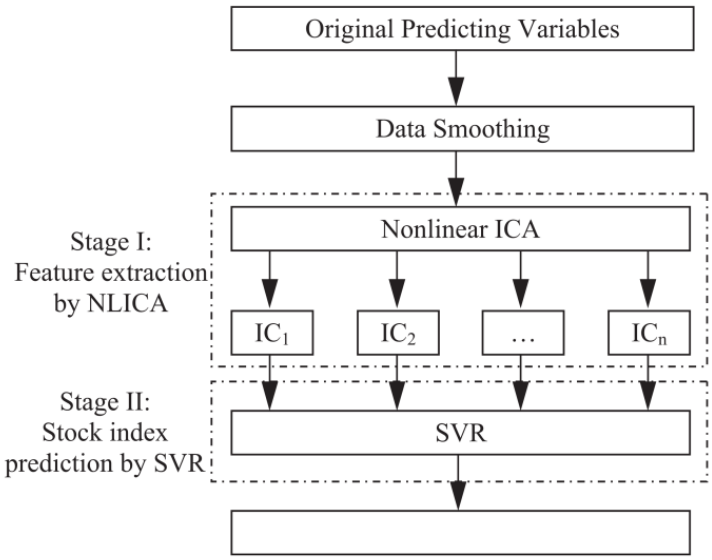
	DS (%)	MSE	RMSE	NMSE	MAE	MAPE
(a) NZD/AUD						
SVR						
2007	**55.02**	3.1903e−005	0.0056	2.0218e−006	0.0037	2.3705e−005
2008	45.28	5.4424e−005	0.0074	3.4149e−006	0.0050	6.9396e−006
cSVR						
2007	53.41	**1.5189e−005**	**0.0039**	**9.6063e−007**	**7.1164e−004**	**1.9383e−006**
2008	**46.40**	**6.5802e−006**	**0.0026**	**4.1534e−007**	**4.8927e−004**	**5.5700e−007**
(b) NZD/EUD						
SVR						
2007	52.61	1.9504e−005	0.0044	1.2360e−006	0.0032	2.1510e−005
2008	45.28	2.6656e−005	0.0052	1.6726e−006	0.0037	6.6451e−005
cSVR						
2007	**53.82**	**9.1822e−008**	**3.0302e−004**	**5.8073e−009**	**1.8954e−004**	**1.1167e−006**
2008	**47.60**	**1.2423e−006**	**0.0011**	**7.8413e−008**	**3.0918e−004**	**1.5854e−006**
(c) NZD/GBP						
SVR						
2007	**54.62**	1.2368e−005	0.0035	7.8380e−007	0.0024	1.9437e−005
2008	**51.57**	2.1833e−005	0.0047	1.3699e−006	0.0032	4.3566e−005
cSVR						
2007	53.82	**2.4687e−006**	**0.0016**	**1.5614e−007**	**3.0980e−004**	**4.4691e−006**
2008	50.00	**3.5398e−006**	**0.0019**	**2.2343e−007**	**2.8227e−004**	**1.4314e−006**
(d) NZD/JPY						
SVR						
2007	**57.83**	1.3234	1.1504	0.0839	0.8159	6.4106e−006
2008	**46.46**	2.0997	1.4490	0.1317	0.9733	9.2072e−005
cSVR						
2007	56.22	**0.0334**	**0.1828**	**0.0021**	**0.0744**	**3.0934e−006**
2008	46.00	**0.6052**	**0.7779**	**0.0382**	**0.2241**	**7.7199e−006**
(e) NZD/USD						
SVR						
2007	55.42	7.3222e−005	0.0086	4.6403e−006	0.0057	1.2783e−005
2008	**47.64**	9.1267e−005	0.0096	5.7266e−006	0.0069	4.0427e−005
cSVR						
2007	**56.22**	**3.0260e−005**	**0.0055**	**1.9138e−006**	**0.0012**	**1.7798e−006**
2008	**46.00**	**7.7186e−006**	**0.0028**	**4.8719e−007**	**7.9949e−004**	**2.4130e−004**

همان‌طور که در جدول۱-۶، مقایسه‌ی تأثیر ورود همبستگی به‌عنوان ورودی آمده است، نتایج در بیشتر ارزهای موردبررسی بهبود پیداکرده است .

ایده‌ی معرفی‌شده در این مقاله تأثیرگذار بوده و می‌تواند یک ویژگی یا نشانگر بازار مناسب به‌عنوان ورودی مدنظر قرار گیرد.

۵-۱ ترکیب آنالیز اجزاء مستقل غیرخطی و ماشین بردار برازشی به‌منظور پیش‌بینی قیمت بورس

در مرجع [۱۴] به دنبال استخراج ویژگی، از قیمت به‌منظور پیش‌بینی بورس است، برای این منظور از روش ICA[1] استفاده و به‌عنوان ورودی به SVR استفاده نموده است.



شکل۱-۶، ساختار پیشنهادی NLICA-SVR

همان‌طور که در شکل۱-۶، ساختار پیشنهادی NLICA-SVR آمده داده‌های قیمت به Nonlinear ICA داده‌شده و IC[2] های خروجی از این روش را به‌عنوان ورودی SVR در نظر گرفته است. معیارهای مقایسه استفاده‌شده و نحوه محاسبه آن‌ها در منبع ذکرشده [۱۴]آمده است.

[1] independent component analysis
[2] Independent component

جهت انجام آزمایش از داده‌ی سهام Shanghai Stock Exchange Composite (SSEC) index و Nikkei 225 stock index استفاده نموده است. روش پیدا نمودن مقادیر بهینه برای SVR نیز روش Grid Search بوده است.

در بخش نتایج روی دو سهام و با روش‌های متفاوت استخراج ویژگی آزمایش‌هایی انجام‌شده است که در ادامه آمده‌اند.

جدول۱-۷، مقایسه‌ی روش پیشنهادی با روش‌های دیگر استخراج ویژگی

T
L

Metrics					
Models	RMSE	MAD	MAPE (%)	RMSPE (%)	DS (%)
NLICA–SVR	**32.20**	**24.94**	**0.16**	**0.21**	**83.69**
LICA–SVR	180.87	130.03	0.85	1.25	68.25
PCA–SVR	353.48	227.59	1.50	2.44	64.41
Single SVR	182.56	134.64	0.88	1.26	68.22

جدول۱-۸، مقایسه‌ی روش پیشنهادی با روش‌های دیگر استخراج ویژگی

Metrics					
Models	RMSE	MAD	MAPE (%)	RMSPE (%)	DS (%)
NLICA–SVR	**3.95**	**2.89**	**2.18**	**2.97**	**71.49**
LICA–SVR	3.97	2.95	2.20	2.99	67.76
PCA–SVR	4.34	3.14	2.38	3.27	60.31
Single SVR	4.35	3.15	2.38	3.28	65.90

نتایج روی سهام Nikkei 225 و SSEC نشان از کاهش خطا با استفاده از روش NLICA دارد که خطای کمتر از روش خطی ICA و نیز روش کاهش بعد PCA به همراه داشته است.

همچنین روش پیشنهادی را با SVR های متفاوت و روش‌های دیگر تخمین تابع موردبررسی قرار داده است که در جدول ۱-۹، مقایسه روش پیشنهادی با مقاله‌های دیگر آمده است.

جدول۱-۹، مقایسه‌ی روش پیشنهادی با مقاله‌های قبلی

Metrics					
Models	RMSE	MAD	MAPE (%)	RMSPE (%)	DS (%)
NLICA–SVR	**32.20**	**24.94**	**0.16**	**0.21**	**83.69**
NLICA–SVR-3steps [32]	38.47	30.34	0.22	0.28	78.36
NLICA–BPN [33]	37.45	28.97	0.21	0.28	79.22
NLICA–SVR–PSO [34]	35.84	26.51	0.19	0.25	80.65

جدول۱-۱۰، مقایسه‌ی روش پیشنهادی با روش مقاله‌های قبلی

Metrics					
Models	RMSE	MAD	MAPE (%)	RMSPE (%)	DS (%)
NLICA–SVR	**3.93**	**2.89**	**2.18**	**2.97**	**71.49**
NLICA–SVR-3steps [32]	4.08	3.06	2.31	3.16	64.62
NLICA–BPN [33]	3.98	2.99	2.21	2.99	66.95
NLICA–SVR–PSO [34]	3.95	2.94	2.20	2.97	70.04

که در مقایسه با ۳ روش دیگر روش بهینه‌سازی SVR با استفاده از Grid Search جواب بهتری به همراه داشته و روش بهینه‌سازی SVR-PSO در رتبه دوم با اختلاف کمی قرارگرفته است.

۱-۵-۱ نقد ترکیب آنالیز اجزاء مستقل غیرخطی و ماشین بردار برازشی به‌منظور پیش‌بینی قیمت بورس

در این مقاله نیز ورودی جدیدی معرفی‌شده که تأثیر خوبی در بحث پیش‌بینی قیمت داشته، ولی ویژگی‌های دیگر یعنی نشانگر بازارها را مدنظر قرار نداده درحالی‌که بیشتر معامله گران بورس به آن‌ها نیز در بحث پیش‌بینی دقت می‌کنند

۱-۶ پیش‌بینی روند بورس بر پایه‌ی روش انتخاب ویژگی فرکتال و ماشین بردار پشتیبان

هدف این مقاله پیش‌بینی روند سهام با استفاده از یک سری ویژگی نشانگر بازار است. بدین شکل که فردا

را پیش‌بینی کند که آیا صعودی است و یا نزولی.

برای شناسایی میزان صحت روش پیشنهادی از داده‌های سهام شانگهای SSECI[1] استفاده کرده و ۱۹ نشانگر بازار از قیمت استخراج کرده که به همراه نحوه محاسبه، در مرجع [۲] آمده است.

به این خاطر که هر نشانگر بازار مقادیر متفاوتی داشته و می‌خواهیم همه‌ی مقادیر را در یک بازه داشته باشیم که بتوان به‌صورت بهتری عمل دسته‌بندی را انجام دهیم با استفاده از فرمول زیر

$$x_i = \frac{x_i - \bar{x}_i}{\sigma_i} \qquad (۲-۱)$$

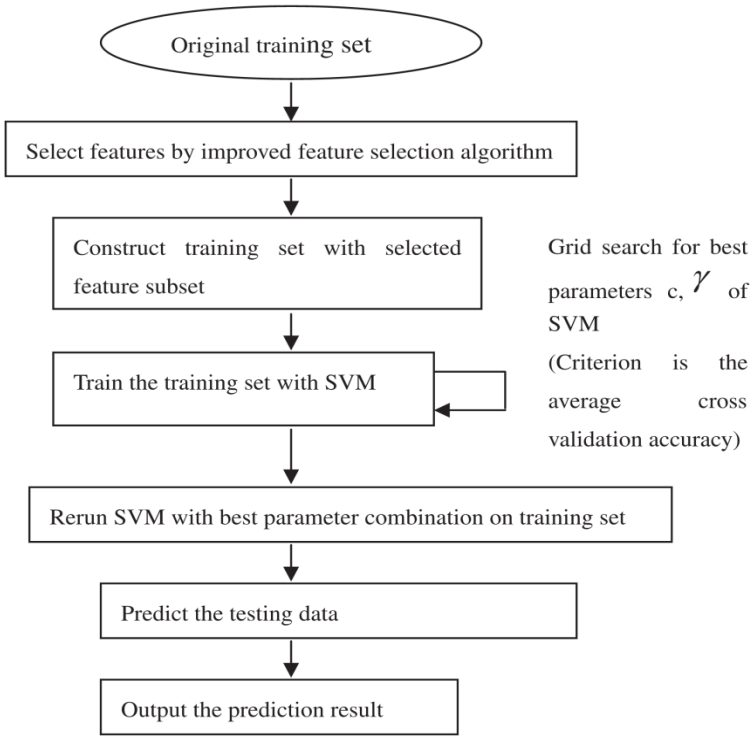
اعداد را نرملایز کردیم و در بازه‌ی ۱ تا -۱ قراردادیم. که در آن $\bar{x}_i$ نشان‌دهنده‌ی میانگین و σ_i نشان‌دهنده‌ی واریانس داده‌ها است.

به‌منظور انتخاب ویژگی از بین ویژگی‌های موجود روش انتخاب ویژگی فرکتالی را پیشنهاد کرده است، که با استفاده از بعد فرکتالی که در زیر آمده بعد کل داده‌ها را محاسبه کرده است و سپس با استفاده از الگوریتم بهینه‌سازی مورچگان به دنبال تعداد ویژگی گشته که بتوانند با تعداد متغیر کمتری نزدیک‌ترین مقدار را به این میزان داشته باشند پیدا کند.

$$D_q = \begin{cases} \lim\limits_{r\to 0} \frac{\sum_i p_i \log p_i^q}{\log r} & q = 1 \\ \lim\limits_{r\to 0} \frac{1}{q-1} \frac{\log \sum_i p_i^q}{\log r} & q \neq 1 \end{cases} \qquad r \in [r_1, r_2] \qquad (۳-۱)$$

درنهایت ویژگی‌های انتخاب‌شده به‌عنوان ورودی به SVM داده‌شده و مقادیر SVM توسط جستجوی Grid پیداشده‌اند.

[1] Shanghai Stock Exchange Composite index



شکل۱-۷، ساختار کلی پیش‌بینی

با استفاده از داده‌ی سهام SSECI[1] ویژگی‌های موردنظر را استخراج کرده و نتایج را با روش‌های دیگر استخراج ویژگی مورد مقایسه قرار داده است.

جدول۱-۱۱، مقایسه‌ی روش‌های انتخاب ویژگی و نتایج هر یک

T

Feature selection methods	Selected feature subset	Prediction accuracy (%)
SVM + Information Gain	MTM, MFI	52.479
SVM + Symmetrical uncertain	MTM, MFI	52.479
SVM + ReliefF	*K*, TRIX, VR, PSY	46.694
SVM + Cfs	MFI	61.984
SVM + OneR	VR, PDI, RSI, ROC	64.463
SVM + improved fractal feature selection	AR, RSI, VRSI, MFI	64.050
SVM	All the features	62.810

[1] Stock Exchange Composite Index

همان‌طور که در جدول۱-۱۱، مقایسه‌ی روش‌های انتخاب ویژگی و نتایج هر یک نشان داده‌شده است، روش انتخاب ویژگی فرکتالی توانسته بین بهترین صحت‌های پیش‌بینی قرار گیرد.

جدول۱-۱۲، مقایسه‌ی روش‌های انتخاب ویژگی و نتایج هر یک

1

Feature selection methods	Selected feature subset	Prediction accuracy (%)
SVM + Information Gain	MFI, CCI, *D*, VR, CR, TRIX, AR, ROC, MTM, PSY	53.252
SVM + Symmetrical uncertain	MFI, *D*, VR, TRIX, CCI, CR, AR, ROC, MTM, PSY	53.252
SVM + ReliefF	BOLL, MFI, VR	45.935
SVM + Cfs	*D*, VR, AR, TRIX, CCI, MFI	51.220
SVM + OneR	MFI, RSI, MTM	50.407
SVM + improved fractal feature selection	MDI, VRSI, MFI	53.659
SVM	All the features	44.716

و حتی در مجموعه داده‌ی دوم به صحت بهترین نسبت به روش‌های دیگر انتخاب ویژگی دست پیداکرده است.

همان‌طور که در نتایج مشخص است انتخاب ویژگی نسبت به استفاده از کل داده‌ها، تأثیر بسیار خوبی در صحت دسته‌بندی داشته است.

در این پژوهش روش انتخاب ویژگی به‌صورت واضح بیان‌نشده، در بخش نتایج نیز فقط معیار صحت استفاده‌شده است و حرفی از انحراف معیار به میان نیامده است و نیز نتایج با استفاده از روش اعتبار سنجی 2-fold محاسبه‌شده که نمی‌تواند به‌صورت خوبی کارایی کلی روش را بیان کند.

۱-۷ بهبود روش کمترین مربعات ماشین بردار پشتیبان به‌منظور پیش‌بینی روند بورس

در این مقاله به بهبود روش LSSVM[1] پرداخته است.[۲۲]

[1] Least Square Support vector machine

برای بهبود کارایی روش ماشین بردار پشتیبان از ترکیب سه هسته[1] استفاده نموده است.

$$\begin{cases} K = \lambda_1 K_{\text{poly}} + \lambda_2 K_{\text{rbf}} + \lambda_3 K_{\text{sig}} \\ \text{where} \quad \lambda_1 + \lambda_2 + \lambda_3 = 1 \\ \qquad\quad 0 \le \lambda_1, \lambda_2, \lambda_3 \le 1. \end{cases} \qquad (۴-۱)$$

همان‌طور که در فرمول فوق آمده است، به‌جای یک هسته از ترکیب ۳ هسته استفاده‌شده است، که باید قانون مرسر[2] را ارضا کنند. که این هسته‌ها شامل هسته‌ی چندجمله‌ای[3]، شعاعی[4] و حلقوی[5] می‌باشند، مقادیر وزن مربوط به هر هسته توسط الگوریتم ژنتیک و با تابع هدف زیر محاسبه می‌شوند.

$$f = E_{\text{accuracy}} - \alpha E_{\text{complexity}} \qquad (۵-۱)$$

که در فرمول فوق صحت و معیار پیچیدگی در ادامه آمده‌اند.

$$E_{\text{accuracy}} = \frac{\text{The number of correct classification}}{\text{The number of all evaluation sample}}. \qquad (۶-۱)$$

$$E_{\text{complexity}} = \frac{n_v}{N_{\text{tot}}} \qquad (۷-۱)$$

که در فرمول پیچیدگی Nv تعداد متغیرهای استفاده‌شده و Ntot تعداد کل متغیرها می‌باشد.

مجموعه داده‌ی استفاده‌شده از دانشگاه پیلسینوانیا[6] گرفته‌شده است، که شامل سهام S&P Index, DJIA NYSE index می‌باشد که داده‌های ماهانه را شامل می‌شوند. خروجی مدنظر نیز صعودی یا نزولی بودن ماه آینده را تشکیل می‌دهد.

به‌منظور مقایسه روش با روش‌های دیگر روی‌داده‌ی آزمایش معیار Hit ratio مورداستفاده قرار گرفت.

---

[1] Kernel
[2] Mercer
[3] polynomial
[4] RBF
[5] sigmoid
[6] Wharton Research Data Service (WRDS), provided by Wharton School,University of Pennsylvania, Philadelphia

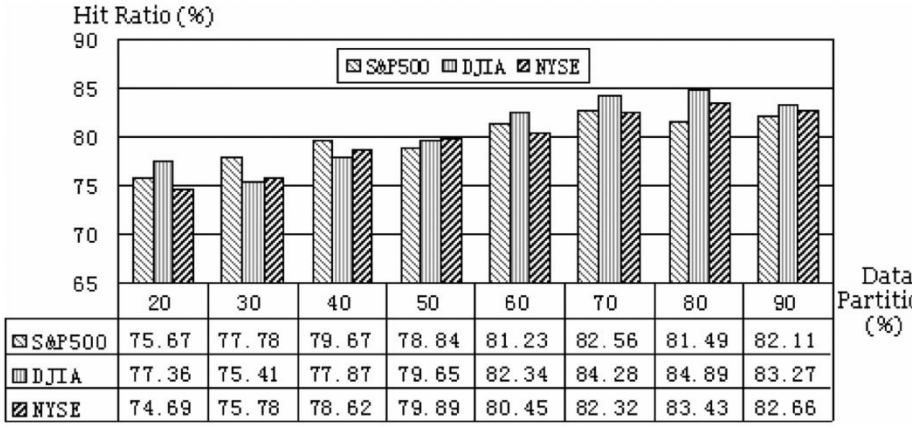
$$\text{Hit ratio} = \frac{1}{N}\sum_{i=1}^{N} R_i \qquad (۸-۱)$$

که نشان‌دهنده‌ی میزان صحت درستی پیش‌بینی است، که در آن Ri، درصورتی‌که مقدار پیش‌بینی با مقدار واقعی برابر باشد مقدار ۱ به خود گرفته و در غیر این صورت مقدار صفر می‌گیرد. درنتیجه این معیار درصد پیش‌بینی درست روش را نشان می‌دهد. ویژگی‌های مورداستفاده به تعداد۲۶تا به‌عنوان ورودی در مرجع ذکرشده آمده است. که شماره‌های ۱ تا ۱۸ را ورودی نشانگر بازار تشکیل داده و ۱۹ تا ۲۶ را اخبار تشکیل می‌دهند.

۱-۷-۱ نتایج

SELECTED KEY FEATURES BY GA-BASED FEATURE SELECTION PROCEDURE

Data Partition (*i*%)	Selected Feature ID		
	S&P 500	DJIA	NYSE
20	1, 7, 11, 17, 19, 20, 21, 26	1, 8, 11, 20, 21, 23	2, 7, 16, 20, 22, 24, 25
30	2, 5, 11, 19, 20, 21	7, 19, 20, 21, 23	1, 15, 20, 21, 22, 25
40	11, 16, 19, 20, 21	3, 15, 19, 20, 23	4, 9, 20, 22, 24
50	1, 5, 13, 15, 20, 21	1, 20, 21, 23, 26	12, 18, 20, 22, 25
60	1, 11, 17, 20, 21	1, 10, 17, 20, 23	15, 19, 20, 22, 24
70	8, 11, 20, 21, 26	11, 16, 19, 20, 23	1, 19, 20, 22, 24
80	7, 15, 19, 20, 22, 26	1, 19, 20, 23	11, 16, 20, 22, 25
90	11, 19, 20, 21	9, 20, 21, 23, 26	1, 20, 22, 24, 25



	20	30	40	50	60	70	80	90
S&P500	75.67	77.78	79.67	78.84	81.23	82.56	81.49	82.11
DJIA	77.36	75.41	77.87	79.65	82.34	84.28	84.89	83.27
NYSE	74.69	75.78	78.62	79.89	80.45	82.32	83.43	82.66

شکل۱-۸ مقایسه نتایج با ویژگی‌ها و درصد داده‌ی استفاده‌شده برای یافتن آن‌ها

به‌منظور نشان دادن کارایی روش، آزمایش را با داده‌های آزمایش متفاوت موردبررسی قراردادِ است. بدین شکل که به‌صورت تصادفی یک‌بار ۲۰ درصد داده‌ها را به‌منظور آموزش استفاده کرده و بقیه را به‌عنوان آزمون و این کار را تا ۹۰ درصد به‌عنوان داده‌ی آموزش تکرار کرده، نتایج حاصل از انتخاب ویژگی با درصد داده‌ها متفاوت تفاوت‌هایی داشت که می‌توان با نگاه کردن به ویژگی‌های مشترک در بین هر هشت مورد به ویژگی‌های مناسب برای پیش‌بینی آن سهام پی برد. همان‌طور که در نتایج شکل۱- مشخص است نتایج در استفاده از ۹۰ درصد داده‌ها با کاهش مواجه شده است. نکته‌ی قابل‌توجه در ویژگی‌های استخراج‌شده این است که بیشتر آن‌ها را اخبار تشکیل می‌دهند و این نشانه این است که بازار بیشتر با اخبار حرکت می‌کند تا ویژگی‌های نشانگر بازار.

جدول۱-۱۳، مقایسه‌ی روش‌های جستجوی متفاوت

Index	Parameter search method	Prediction performance (%)	CPU time (s)
S&P 500	Grid Search	82.81	14289.40
	Direct Search	82.59	1568.92
	Genetic Algorithm	82.66	2539.23
DJIA	Grid Search	85.57	15324.17
	Direct Search	85.33	1624.84
	Genetic Algorithm	85.28	2618.56
NYSE	Grid Search	84.66	14788.38
	Direct Search	84.02	1598.45
	Genetic Algorithm	84.13	2589.62

با توجه به جدول۱-۱۳، به‌منظور محاسبه‌ی مقادیر ماشین بردار پشتیبان بهترین کارایی را الگوریتم Grid از خود نشان داده است ولی زمانی چندین برابر روش‌های مستقیم و الگوریتم ژنتیک صرف کرده است، که دلیل آن بررسی همه‌ی حالات ممکن می‌باشد.

جدول۱-۱۴، مقایسه‌ی کلیه‌ی روش‌ها باهدف پیش‌بینی

Model	Hit ratio (%)		
	S&P 500	DJIA	NYSE
ARIMA	55.78 [6.45]	56.74 [6.76]	56.32 [5.98]
LDA	61.43 [5.85]	59.33 [6.51]	60.28 [7.82]
BPNN	67.56 [8.34]	65.44 [7.85]	66.87 [8.43]
Standard SVM	72.61 [5.62]	71.58 [6.36]	70.84 [7.05]
$LSSVM_{poly}$	70.43 [6.33]	70.47 [5.89]	69.39 [6.56]
$LSSVM_{rbf}$	75.45 [6.76]	74.84 [6.12]	73.43 [5.87]
$LSSVM_{sig}$	74.78 [7.12]	74.13 [6.66]	74.25 [7.82]
$LSSVM_{mix}$	77.86 [6.01]	79.22 [7.35]	78.74 [7.19]
$LSSVM_{gafs}$	82.43 [4.54]	84.87 [6.58]	83.69 [5.45]
$LSSVM_{gapo}$	82.58 [5.03]	84.92 [7.91]	83.97 [7.56]
Evolving LSSVM	82.66 [5.17]	85.28 [5.85]	84.13 [6.02]

* $LSSVM_{poly}$: individual LSSVM model with polynomial kernel; $LSSVM_{rbf}$: individual LSSVM model with RBF kernel; $LSSVM_{sig}$: individual LSSVM model with sigmoid kernel; $LSSVM_{mix}$: individual LSSVM model with a mixed kernel; $LSSVM_{gafs}$: individual LSSVM model with GA-based input feature selection only; $LSSVM_{gapo}$: individual LSSVM model with GA-based parameter optimization only.

در جدول ۱-۱۴، به مقایسه روش پیشنهادی با روش‌های تک هسته و بدون استفاده از انتخاب ویژگی و روش‌های قدیمی پرداخته که روش پیشنهادی به بهترین میزان صحت دست پیداکرده است.

در این پژوهش از ویژگی‌های بسیار مناسبی استفاده‌شده است، ولی ایراد آن این است که داده‌ها به‌صورت عمومی در دسترس نبوده و نمی‌توان از داده‌های اقتصادی استفاده‌شده در این پژوهش استفاده نمود. و همچنین انتخاب ویژگی آن تکاملی است و در هر اجرا به ویژگی‌های متفاوتی دست پیدا می‌کند.

۱-۸ نتیجه‌گیری

در این بخش به مقایسه‌ی پژوهش‌های مختلفی که درزمینه‌ی پیش‌بینی بورس بیان‌شده‌اند پرداختیم بدین شکل که مقاله‌ها را به دودسته بر اساس هدف پیش‌بینی آن‌ها تقسیم می‌کردیم یعنی پیش‌بینی روند و قیمت، مقاله‌های پیش‌بینی روند به مقاله‌هایی گفته می‌شود که به دنبال دسته‌بندی روند بازار هستند و کار آن‌ها پیدا کردن روند بازار است که صعودی است و یا نزولی، دسته‌ی دوم پا را فراتر گذاشته و به دنبال پیش‌بینی قیمت بازار هستند.

در این مقاله‌ها از داده‌ی متفاوتی استفاده‌شده است اما معمولاً داده‌ها روزانه بوده، اما داده‌ی ماهانه نیز در آن‌ها به چشم می‌خورد.

اکثر روش‌های استفاده‌شده در این مقاله‌ها را ماشین بردار پشتیبان تشکیل می‌دهد که به دو شکل SVM[1] و SVR[2] آورده شده است که SVM به‌منظور دسته‌بندی داده‌ها در دودسته‌ی صعود و یا نزول قیمت استفاده می‌شود و SVR به‌منظور تخمین قیمت به کار می‌رود.

به‌منظور پیش‌بینی قیمت بیشتر از قیمت به همراه ویژگی‌های استخراج‌شده از آن استفاده‌شده و برای پیش‌بینی روند از نشانگر بازارها بیشتر استفاده‌شده است و در برخی موارد از اخبار نیز استفاده‌شده است که تأثیر اخبار به‌مراتب بیشتر از ویژگی‌های نشانگر بازار بوده است.

در بیشتر کارها از قیمت Log گرفته‌شده و نیز در بیشتر کارها از نرمال کردن Z-score استفاده‌شده است که با استفاده از میانگین و واریانس داده‌ها را به فضای بین [1,1-] می‌برد.

طول پنجره دردوکار ۳ و ۴ بوده ولی در یک کار ۲۰ نیز بوده است که نشان‌دهنده‌ی متغیر بودن طول آن می‌باشد.

در بحث بهینه‌سازی ماشین بردار پشتیبان بهترین صحت را روش Grid از خود نشان داده است ولی زمان آن نسبت به روش‌های تکاملی بسیار بیشتر است چون کلیه‌ی حالات ممکن را موردبررسی قرار می‌دهد.

به‌منظور پیدا نمودن بهترین صحت بیشتر کارها از روش اعتبار سنجی 5-fold استفاده کرده‌اند که کار معقولی به نظر می‌رسد که در بیان دیگر همان ۲۰ درصد برای آزمون و ۸۰ درصد برای آموزش را نشان می‌دهد.

[1] Support vector machine
[2] Support vector regression

در مقاله‌های پیش‌بینی روند همه از معیار Accuracy یا به اسم دیگر Hit Ratio استفاده کرده بودند ولی حرفی از انحراف معیار و دیگر معیارها در بحث دسته‌بندی دودسته زده نشده بود که ایرادی بر کار آن‌ها می‌باشد.

در برخی کارها از روش‌های تکاملی به‌منظور انتخاب ویژگی استفاده‌شده بود که این کار باعث می‌شود که به ویژگی‌های متفاوتی در اجراهای متفاوت دست پیدا کنیم.

همچنین در هر مقاله از یک سری ورودی خاص استفاده‌شده بود و به بقیه‌ی ورودی‌ها ممکن پرداخته نشده بود.

۱-۹ ساختار پایان‌نامه

فصل اول

کلیات تحقیق : در این فصل مقدمه‌ای از اهمیت پیش‌بینی سهام در بورس و همچنین کارهایی که در این راستا انجام‌گرفته آورده شده است و درنهایت نتیجه‌گیری از کل فصل و نقد روش‌ها بیان‌شده است.

فصل دوم

کاهش ویژگی و کلاسه‌بندی : این فصل به بررسی روش ماشین بردار پشتیبان و همچنین کاربردهای SVM و دلایل استفاده از این روش اشاره‌شده است.

نحوه کارکرد ماشین بردار پشتیبان خطی و محاسباتی که در این روش به کار می‌رود و خلاصه استفاده عملی از SVM به‌طور کامل بحث شده است.

فصل سوم

اندیکاتورها : شناسایی و بررسی نشانگرهایی که در تحقیق مورداستفاده پیاده‌سازی شده و همچنین تعریف آن‌ها در این فصل همراه با فرمول‌های محاسباتی هرکدام آورده شده است.

فصل چهارم

الگوریتم پیشنهادی : روش پیشنهادی جهت بهبود و تسریع درروند پیش‌بینی در این فصل با کمک فلوچارت‌ها و توضیح مرحله‌ای که در پیاده‌سازی انجام‌گرفته تشریح شده است.

ضمایم

درنهایت نرم‌افزارهایی که جهت پیاده‌سازی در این تحقیق اعم از دریافت داده‌های سهام و تحلیل و تفکیک داده‌ها ، کلاسه‌بندی مقایسه و آزمودن استفاده‌شده به‌طور مختصر در بخش ضمایم آورده شده است.

فصل دو

مروری بر منابع و ادبیات موضوع

(کاهش ویژگی و طبقه بندی)

مقدمه

پیشرفت‌های به وجود آمده در جمع‌آوری داده و قابلیت‌های ذخیره‌سازی در طی دهه‌های اخیر باعث شده در بسیاری از علوم با حجم بزرگی از اطلاعات روبرو شویم. محققان در زمینه‌های مختلف مانند مهندسی، ستاره‌شناسی، زیست‌شناسی و اقتصاد هرروز با مشاهدات بیشتر و بیشتری روبرو می‌شوند. روش‌های آماری سنتی به دو دلیل امروزه کارایی خود را ازدست‌داده‌اند. علت اول افزایش تعداد مشاهدات[1] است، و علت دوم که از اهمیت بالاتری برخوردار است افزایش تعداد متغیرهای مربوط به یک مشاهده می‌باشد.

در علوم کامپیوتر و یادگیری ماشین بیشتر از عبارات "ویژگی"[2] و یا "صفت"[3] برای تعداد متغیرها استفاده می‌گردد. که در این پایان‌نامه نیز نشانگرها به‌عنوان ویژگی‌ها به‌حساب می‌آیند. روش‌های متعددی برای انتخاب ویژگی و یا کاهش ویژگی تاکنون مطرح‌شده که نوع داده‌ها و تعداد آن‌ها می‌تواند مشخص‌کننده روش انتخابی باشد. در این فصل به معرفی انواع این روش‌ها خواهیم پرداخت و درنهایت یک جمع‌بندی از کلیه روش‌ها آورده شده است.

۲-۱- روش‌های کاهش ابعاد داده به دودسته تقسیم می‌شوند:

۲-۱-۱- روش‌های مبتنی بر استخراج ویژگی:

این روش‌ها یک فضای چندبعدی را به یک فضای با ابعاد کمتر نگاشت می‌کنند. درواقع با ترکیب مقادیر ویژگی‌ها موجود، تعداد کمتری ویژگی به وجود می‌آورند به‌طوری‌که این ویژگی‌ها دارای تمام (یا بخش اعظمی از) اطلاعات موجود در ویژگی‌ها اولیه باشند. این روش‌ها به دودسته‌ی خطی و غیرخطی

[1] Observations
[2] Feature
[3] Attribute

تقسیم می‌شوند.

از روش‌های خطی می‌توان به DFT، DWT، PCA و FA اشاره کرد که دو تا از آن‌ها را که مرتبط با این تحقیق می‌باشد در ادامه‌ی همین فصل توضیح خواهیم داد. روش‌های دیگر غیرخطی عبارت‌اند از:

Projection Pursuit (PP) : برخلاف روش‌های PCA و FA می‌تواند اطلاعات بالاتر از مرتبه‌ی دوم را ترکیب نماید. بنابراین روش مناسبی برای بسترهای داده‌ای غیر گاوسی است.

Independent Component Analysis (ICA) : این روش نیز یک نگاشت خطی انجام می‌دهد اما بردارهای این نگاشت لزوماً بر یکدیگر عمود نیستند، درحالی‌که در روش‌های دیگر مانند PCA این بردارها بر هم عمودند.

Random Projection (PP) : یک روش ساده و درعین‌حال قدرتمند برای کاهش ابعاد داده است که از ماتریس‌های نگاشت تصادفی برای نگاشت داده‌ها به یک فضای با ابعاد کمتر استفاده می‌کند.

از روش‌های غیرخطی نیز می‌توان به موارد زیر اشاره کرد:

- Principal Curves
- Self Organizing Maps
- Vector Quantization
- Genetic and Evolutionary Algorithms
- Regression

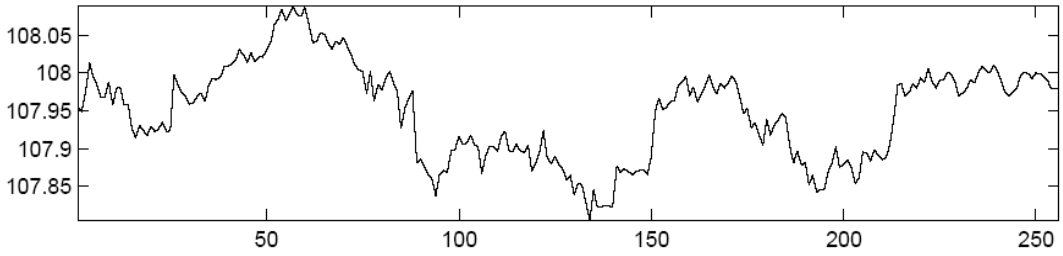
در پژوهش [3] نگاهی اجمالی به کلیه‌ی روش‌های کاهش ابعاد داده‌ی مبتنی بر استخراج ویژگی شده است.. برای تهیه‌ی بیشتر مطالبی که در این دو بخش ارائه‌شده از منبع [4] که یک پایان‌نامه دکتری درزمینه‌ی داده‌کاوی بر روی سری‌های زمانی می‌باشد استفاده‌شده است. روش PCA که بهترین تبدیل خطی به‌حساب می‌آید را بیان خواهیم کرد. برای تهیه‌ی این بخش نیز از منبع [5] استفاده کرده‌ایم که

یک tutorial بسیار عالی می‌باشد. و به همین دلیل به توضیح اجمالی بسنده شده است.

Discrete Fourier Transform (DFT)

در بسیاری از کاربردها مرسوم است که از ترکیب توابع پایه‌ای برای تقریب یک تابع استفاده شود. به‌عنوان‌مثال هر تابع پیوسته را می‌توان توسط مجموعه‌ای از توابع چندجمله‌ای نمایش داد. تبدیل فوریه نوعی تبدیل است که یک تابع را به‌صورت توابع پایه‌ای سینوسی که هرکدام در مقادیری ضرب شده‌اند نشان می‌دهد. از تبدیل فوریه در بسیاری از زمینه‌های علمی مانند فیزیک، هندسه، آمار و پردازش سیگنال استفاده می‌شود.

تبدیل فوریه یک تبدیل برگشت‌پذیر است. این تبدیل می‌تواند به دو صورت پیوسته یا گسسته انجام شود. در کامپیوتر و بخصوص در پردازش سیگنال معمولاً از تبدیل فوریه‌ی گسسته (DFT) استفاده می‌شود. خوشبختانه الگوریتم‌های سریعی تحت عنوان FFT[1] برای تبدیل فوریه‌ی گسسته به وجود آمده است. به‌عنوان‌مثال بردار $\vec{x}$ که سری زمانی مربوط به قیمت روزانه‌ی سهام شرکت IBM در سال ۲۰۰۱ می‌باشد را در نظر بگیرید. این بردار که از ۲۵۶ مؤلفه تشکیل‌شده است را در شکل زیر نشان داده‌ایم.

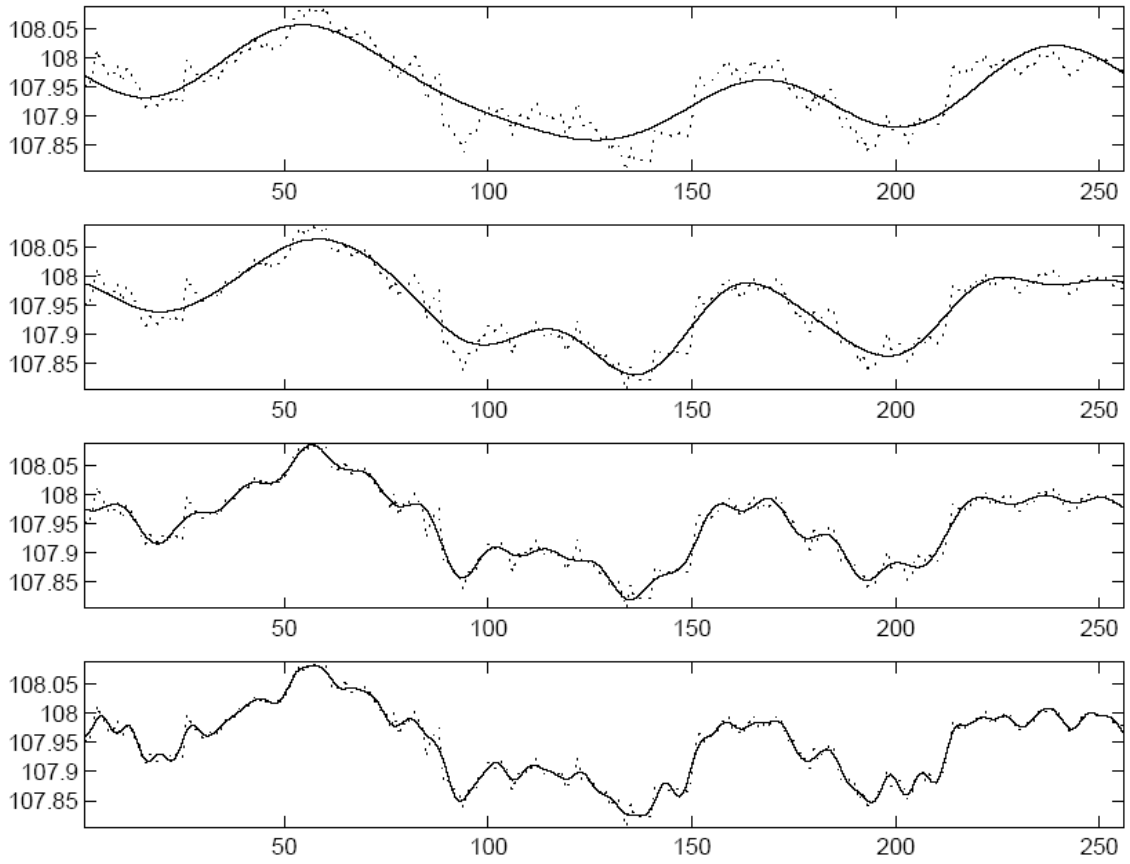


شکل۲-۱ قیمت روزانه سهام شرکت IBM در سال ۲۰۰۱

اکنون بردار $\vec{X} = DFT(\vec{x})$ را محاسبه می‌کنیم. مسلماً $\vec{X}$ نیز دارای ۲۵۶ مؤلفه خواهد بود. اگر فقط ۱۰ ضریب پراهمیت‌تر در $\vec{X}$ را نگه‌داشته و بقیه را حذف نماییم، نموداری شبیه به اولین نمودار در شکل ۲-۲

[1] Fast Fourier Transform

به دست خواهیم آورد. در این شکل مشاهده می‌کنید که هرچه تعداد ضرایب بیشتری را نگه‌داری نماییم، جزئیات نمودار بیشتر حفظ خواهد شد.

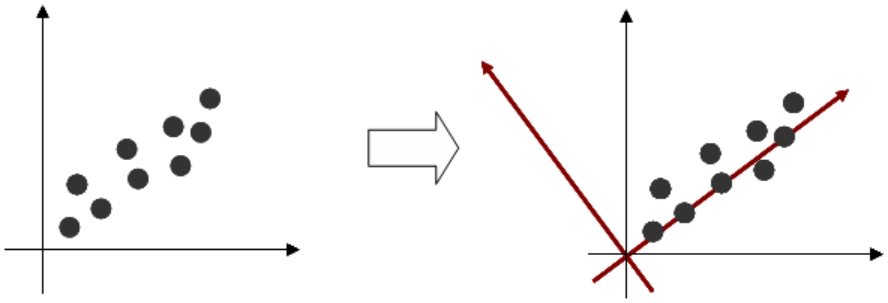


شکل۲-۲ بازسازی نمودار قیمت روزانه سهام شرکت IBM با ضرایب به‌دست‌آمده از DFT این نمودارها از بالا به پایین به ترتیب با حفظ ۱۰، ۲۰، ۴۰ و ۸۰ ضریب DFT به‌دست‌آمده‌اند

Principal Component Analysis (PCA)

فن PCA بهترین روش برای کاهش ابعاد داده به‌صورت خطی می‌باشد. یعنی با حذف ضرایب کم-اهمیت به‌دست‌آمده از این تبدیل، اطلاعات ازدست‌رفته نسبت به روش‌های دیگر کمتر است. البته کاربرد PCA محدود به کاهش ابعاد داده نمی‌شود و در زمینه‌های دیگری مانند شناسایی الگو و تشخیص چهره نیز مورداستفاده قرار می‌گیرد. در این روش محورهای مختصات جدیدی برای داده‌ها تعریف‌شده و داده‌ها بر اساس این محورهای مختصات جدید بیان می‌شوند. اولین محور باید در جهتی قرار گیرد که واریانس

داده‌ها ماکزیمم شود (یعنی در جهتی که پراکندگی داده‌ها بیشتر است). دومین محور باید عمود بر محور اول به‌گونه‌ای قرار گیرد که واریانس داده‌ها ماکزیمم شود. به همین ترتیب محورهای بعدی عمود بر تمامی محورهای قبلی به‌گونه‌ای قرار می‌گیرند که داده‌ها در آن جهت دارای بیشترین پراکندگی باشند. در شکل زیر این مطلب برای داده‌های دوبعدی نشان داده‌شده است.



شکل ۲-۳ انتخاب محورهای جدید برای داده‌های دوبعدی

۲-۱-۲- روش‌های مبتنی بر انتخاب ویژگی:

این روش‌ها سعی می‌کنند با انتخاب زیرمجموعه‌ای از ویژگی‌ها اولیه، ابعاد داده‌ها را کاهش دهند.

مسئله انتخاب ویژگی، یکی از مسائلی است که در مبحث یادگیری ماشین و همچنین شناسایی آماری الگو مطرح است. این مسئله در بسیاری از کاربردها (مانند طبقه‌بندی) اهمیت به سزایی دارد، زیرا در این کاربردها تعداد زیادی ویژگی وجود دارد، که بسیاری از آن‌ها یا بلااستفاده هستند و یا اینکه بار اطلاعاتی چندانی ندارند. حذف نکردن این ویژگی‌ها مشکلی ازلحاظ اطلاعاتی ایجاد نمی‌کند ولی بار محاسباتی را برای کاربرد موردنظر بالا می‌برد. و علاوه بر این باعث می‌شود که اطلاعات غیرمفید زیادی را به همراه داده‌های مفید ذخیره کنیم.

در این بخش ما در ابتدا تعاریفی که برای انتخاب ویژگی ارائه‌شده‌اند و همچنین، تعاریف موردنیاز برای درک این مسئله را ارائه می‌دهیم. سپس روش‌های مختلف برای این مسئله را بر اساس نوع و ترتیب تولید زیرمجموعه ویژگی‌های کاندید و همچنین نحوه ارزیابی این زیرمجموعه‌ها دسته‌بندی می‌کنیم. سپس

تعدادی از روش‌های معرفی‌شده در هر دسته را معرفی و بر اساس اهمیت، تا جایی که مقدور باشد، آن‌ها را تشریح و الگوریتم برخی از آن‌ها را ذکر می‌کنیم. لازم به ذکر است که به دلیل اینکه مبحث انتخاب ویژگی به مبحث طبقه‌بندی بسیار نزدیک است، بعضی از مسائلی که در اینجا مطرح می‌شود مربوط به مبحث طبقه‌بندی می‌باشد.

تعاریف

مسئله انتخاب ویژگی به‌وسیله نویسندگان مختلف، از دیدگاه‌های متفاوتی موردبررسی قرارگرفته است. هر نویسنده نیز با توجه به نوع کاربرد، تعریفی را از آن ارائه داده است. در ادامه چند مورد از این تعاریف را بیان می‌کنیم[6]:

۱. **تعریف ایدئال:** پیدا کردن یک زیرمجموعه با حداقل اندازه ممکن، برای ویژگی‌ها است، که برای هدف موردنظر اطلاعات لازم و کافی را در برداشته باشد. بدیهی است که هدف تمام الگوریتم‌ها و روش‌های انتخاب ویژگی همین زیرمجموعه است.

۲. **تعریف کلاسیک:** انتخاب یک زیرمجموعه M عنصری از میان N ویژگی، به‌طوری‌که $M < N$ باشد و همچنین مقدار یک تابع معیار[1] برای زیرمجموعه موردنظر، نسبت به سایر زیرمجموعه‌های هم‌اندازه دیگر بهینه باشد. این تعریفی است که Fukunaga و Narenda در سال ۱۹۷۷ ارائه داده‌اند.

۳. **افزایش دقت پیشگوئی:** هدف انتخاب ویژگی این است که یک زیرمجموعه از ویژگی‌ها برای افزایش دقت پیشگوئی انتخاب شوند. به‌عبارت‌دیگر کاهش اندازه ساختار بدون کاهش فراوان در دقت پیشگوئی طبقه‌بندی کننده‌ای که با استفاده از ویژگی‌ها داده‌شده به دست می‌آید.

۴. **تخمین توزیع کلاس اصلی:** هدف از انتخاب ویژگی این است که یک زیرمجموعه کوچک از ویژگی-

[1] - Criterion function

ها انتخاب شوند، توزیع ویژگی‌هایی که انتخاب می‌شوند، بایستی تا حد امکان به توزیع کلاس اصلی با توجه به تمام مقادیر ویژگی‌های انتخاب‌شده نزدیک باشد.

روش‌های مختلف انتخاب ویژگی، تلاش می‌کنند تا از میان 2^N زیرمجموعه کاندید، بهترین زیرمجموعه را پیدا کنند. در تمام این روش‌ها بر اساس کاربرد و نوع تعریف، زیرمجموعه‌ای به‌عنوان جواب انتخاب می‌شود، که بتواند مقدار یک تابع ارزیابی را بهینه کند. باوجوداینکه هر روشی سعی می‌کند که بتواند، بهترین ویژگی‌ها را انتخاب کند، اما با توجه به وسعت جواب‌های ممکن، و اینکه این مجموعه‌های جواب به‌صورت توانی با N افزایش پیدا می‌کنند، پیدا کردن جواب بهینه مشکل و در N های متوسط و بزرگ بسیار پرهزینه است.

برای اینکه بتوانیم تقسیم‌بندی درستی از روش‌های مختلف انتخاب ویژگی داشته باشیم، به این صورت عمل می‌کنیم که فرآیند انتخاب ویژگی در تمامی روش‌ها را به این بخش‌ها تقسیم می‌کنیم:

۱. **تابع تولیدکننده**[1]: این تابع زیرمجموعه‌های کاندید را برای روش موردنظر پیدا می‌کند.
۲. **تابع ارزیابی**[2]: زیرمجموعه موردنظر را بر اساس روش داده‌شده، ارزیابی و یک عدد به‌عنوان میزان خوبی روش بازمی‌گرداند. روش‌های مختلف سعی دریافتن زیرمجموعه‌ای دارند که این مقدار را بهینه کند.
۳. **شرط خاتمه:** برای تصمیم‌گیری در مورد زمان توقف الگوریتم.
۴. **تابع تعیین اعتبار**[3]: تصمیم می‌گیرد که آیا زیرمجموعه انتخاب‌شده معتبر است یا خیر؟

تابع تولیدکننده درواقع تابع جستجو است. این تابع زیرمجموعه‌های مختلف را به ترتیب تولید می‌کند، تا

[1] - Generation procedure
[2] - Evaluation function
[3] - Validation procedure

به‌وسیله تابع ارزیابی، مورد ارزیابی قرا بگیرد. تابع تولیدکننده از یکی از حالت‌های زیر شروع به کار می-کند:

۱. بدون ویژگی
۲. با مجموعه تمام ویژگی‌ها
۳. با یک زیرمجموعه تصادفی

در حالت اول ویژگی‌ها به ترتیب به مجموعه اضافه می‌شوند و زیرمجموعه‌های جدید را تولید می‌کنند. این عمل آن‌قدر تکرار می‌شود تا به زیرمجموعه موردنظر برسیم. به این‌گونه روش‌ها، روش‌های پائین به بالا می‌گویند.

در حالت دوم از یک مجموعه شامل تمام ویژگی‌ها، شروع می‌کنیم و به‌مرور و در طی اجرای الگوریتم، ویژگی‌ها را حذف می‌کنیم، تا به زیرمجموعه دلخواه برسیم. روش‌هایی که به این صورت عمل می‌کنند، روش‌های بالا به پائین نام دارند. که روش پیشنهادی در این پایان‌نامه یک نوع از همین روش است

یک تابع ارزیابی، میزان خوب بودن یک زیرمجموعه تولیدشده را بررسی کرده و یک مقدار به‌عنوان میزان خوب بودن زیرمجموعه موردنظر بازمی‌گرداند. این مقدار با بهترین زیرمجموعه قبلی مقایسه می‌شود. اگر زیرمجموعه جدید، بهتر از زیرمجموعه‌های قدیمی باشد، زیرمجموعه جدید به‌عنوان زیرمجموعه بهینه، جایگزین قبلی می‌شود.

باید توجه داشت که بدون داشتن یک شرط خاتمه مناسب، فرآیند انتخاب ویژگی ممکن است، برای همیشه درون فضای جستجو، برای یافتن جواب سرگردان بماند. شرط خاتمه می‌تواند بر پایه تابع تولیدکننده باشد، مانند:

هر زمان که تعداد مشخصی ویژگی انتخاب شدند، هر زمان که به تعداد مشخصی تکرار رسیدیم، و یا اینکه

بر اساس تابع ارزیابی انتخاب شود، مانند: وقتی‌که اضافه یا حذف کردن ویژگی، زیرمجموعه بهتری را تولید نکند، وقتی‌که به یک زیرمجموعه بهینه بر اساس تابع ارزیابی برسیم.

تابع تعیین اعتبار جزئی از فرآیند انتخاب ویژگی نیست، اما در عمل بایستی یک زیرمجموعه ویژگی را در شرایط مختلف آزمودن کنیم تا ببینیم که آیا شرایط موردنیاز، برای حل مسئله موردنظر ما را دارد یا نه؟ برای این کار می‌توانیم از داده‌های نمونه‌برداری شده و یا مجموعه داده‌های شبیه‌سازی‌شده استفاده کنیم.

۲-۳- روش‌های مختلف انتخاب ویژگی

در این بخش ابتدا روش‌های مختلف انتخاب ویژگی را بر اساس دو معیار تابع تولیدکننده و تابع ارزیابی طبقه‌بندی می‌کنیم. سپس آن‌ها را بر اساس عملکرد دسته‌بندی و نحوه اجرای هر دسته را به‌اختصار شرح می‌دهیم.

۲-۳-۱- توابع تولیدکننده

اگر تعداد کل ویژگی‌ها برابر N باشد، تعداد کل زیرمجموعه‌های ممکن برابر 2^N می‌شود. این تعداد برای N های متوسط هم خیلی زیاد است. بر اساس نحوه جستجو در میان این تعداد زیرمجموعه، روش‌های مختلف انتخاب ویژگی را می‌توان به سه دسته زیر تقسیم‌بندی نمود:

۱. جستجوی کامل
۲. جستجوی مکاشفه‌ای
۳. جستجوی تصادفی

در ادامه به معرفی هرکدام از این دسته‌ها می‌پردازیم.

۱ جستجوی کامل[1]

در روش‌هایی که از این نوع جستجو استفاده می‌کنند، تابع تولیدکننده بر اساس تابع ارزیابی استفاده‌شده، تمام فضای جواب (زیرمجموعه‌های ممکن) را برای یافتن جواب بهینه جستجو می‌کند. البته Schlimmer استدلال آورده است که[7]: "کامل بودن جستجو به این معنی نیست که جستجو باید جامع باشد".

توابع مکاشفه‌ای مختلف زیادی طراحی‌شده‌اند، تا جستجو را بدون از دست دادن شانس پیدا کردن جواب بهینه، کاهش دهند. اما با توجه به بزرگی فضای جستجو، $O(2^N)$، این روش‌ها باعث می‌شوند که فضای کمتری جستجو شود. روش‌ها و فن‌های مختلفی برای این کار استفاده‌شده‌اند، بعضی از آن‌ها از فن بازگشت به عقب (Backtracking) نیز در جریان کار استفاده کرده‌اند، مانند: branch and bound، best first search و beam search.

۲ جستجوی مکاشفه‌ای[2]

در روش‌های با این نوع جستجو، در هر بار اجرای الگوریتم، یک ویژگی به مجموعه ویژگی انتخاب‌شده، اضافه و یا از آن حذف می‌شود. به همین دلیل پیچیدگی زمانی آن‌ها محدود و کمتر از $O(N^2)$ می‌باشد. در این‌گونه موارد، اجرای الگوریتم خیلی سریع می‌باشد و پیاده‌سازی آن‌ها نیز بسیار ساده است.

۳ جستجوی تصادفی[3]

روش‌هایی که از این نوع جستجو استفاده می‌کنند، محدوده کمتری از فضای کل حالات را جستجو می‌کنند، که اندازه این محدوده به حداکثر تعداد تکرار الگوریتم بستگی دارد. در این روش‌ها پیدا شدن

---

[1] - Complete Search
[2] - Heuristic Search
[3] - Random Search

جواب بهینه به‌اندازه منابع موجود و زمان اجرای الگوریتم بستگی دارد. در هر بار تکرار، تابع تولیدکننده تعدادی از زیرمجموعه‌های ممکن از فضای جستجو را به‌صورت تصادفی انتخاب می‌کند و در اختیار تابع ارزیابی قرار می‌دهد. تابع تولیدکننده تصادفی پارامترهایی دارد که بایستی تنظیم شوند، تنظیم مناسب این پارامترها در سرعت رسیدن به جواب و پیدا شدن جواب‌های بهتر مؤثر است.

۲-۴- تابع ارزیابی

پیدا شدن یک زیرمجموعه بهینه از مجموعه ویژگی‌ها، به‌صورت مستقیم با انتخاب تابع ارزیابی بستگی دارد. چراکه اگر تابع ارزیابی به زیرمجموعه ویژگی بهینه یک مقدار نامناسب نسبت دهد، این زیرمجموعه هیچ‌گاه به‌عنوان زیرمجموعه بهینه انتخاب نمی‌شود. مقادیری که توابع ارزیابی مختلف به یک زیرمجموعه می‌دهند، باهم متفاوت است.

توابع ارزیابی را می‌توان به طرق مختلفی دسته‌بندی کرد. در اینجا ما دسته‌بندی‌ای که توسط Dash و Liu ارائه‌شده است [6] را بیان می‌کنیم. آن‌ها این معیارها را به پنج دست تقسیم کرده‌اند:

۱. **معیارهای مبتنی بر فاصله**[1]: در این معیارها، مثلاً برای یک مسئله دو کلاسه، یک ویژگی یا یک مجموعه ویژگی مثل X بر یک ویژگی یا یک مجموعه ویژگی دیگر مثل Y ارجحیت دارد، اگر که با آن مجموعه ویژگی مقادیر بزرگ‌تری برای اختلاف بین احتمالات شرطی دو کلاس داشته باشیم. نمونه‌ای از این معیارها همان معیار فاصله اقلیدسی می‌باشد.

۲. **معیارهای مبتنی بر اطلاعات**[2] : این معیارها میزان اطلاعاتی را که به‌وسیله یک ویژگی به دست می‌آید را در نظر می‌گیرند. ویژگی X در این روش‌ها بر ویژگی Y اولویت دارد، اگر اطلاعات به‌دست‌آمده از ویژگی X بیشتر از اطلاعاتی باشد، که از ویژگی Y به دست می‌آید. نمونه‌ای از این

[1] - Distance Measures
[2] - Information Measures

معیارها همان معیار آنتروپی می‌باشد.

۳. **معیارهای مبتنی بر وابستگی**[۱]: این معیارها که با عنوان معیارهای همبستگی[۲] نیز شناخته می‌شوند، قابلیت پیشگوئی مقدار یک متغیر به‌وسیله یک متغیر دیگر را اندازه‌گیری می‌کنند. ضریب (Coefficient) یکی از معیارهای وابستگی کلاسیک است و می‌توانیم آن‌ها برای یافتن همبستگی بین یک ویژگی و یک کلاس به کار ببریم. اگر همبستگی ویژگی X باکلاس C بیشتر از همبستگی ویژگی Y باکلاس C باشد، در این صورت ویژگی X بر ویژگی Y برتری دارد. با یک تغییر کوچک، می‌توانیم وابستگی یک ویژگی با ویژگی‌های دیگر را اندازه‌گیری کنیم. این مقدار درجه افزونگی این ویژگی را نشان می‌دهد. همه توابع ارزیابی بر پایه معیار وابستگی را می‌توانیم بین دودسته معیارهای مبتنی بر فاصله و اطلاعات تقسیم کنیم. اما به خاطر اینکه این روش‌ها از یک دید دیگر به مسئله نگاه می‌کنند، این کار را انجام نمی‌دهیم.

۴. **معیارهای مبتنی بر سازگاری**[۳]: این معیارها جدیدتر هستند و اخیراً توجه بیشتری به آن‌ها شده است. این معیارها خصوصیات متفاوتی نسبت به سایر معیارها دارند، زیرا که به‌شدت به داده‌های آموزشی تکیه‌دارند و در انتخاب یک زیرمجموعه از ویژگی‌ها تمایل دارند، که مجموعه ویژگی‌های کوچک‌تری را انتخاب کنند. این روش‌ها زیرمجموعه‌های با کمترین اندازه را بر اساس از دست دادن یک مقدار قابل‌قبول سازگاری که توسط کاربر تعیین می‌شود، پیدا می‌کنند.

۵. **معیارهای مبتنی بر خطای طبقه‌بندی کننده**[۴]: روش‌هایی که این نوع از تابع ارزیابی را استفاده می‌کنند، با عنوان "wrapper methods" شناخته می‌شوند. دقت عملکرد در این روش‌ها برای تعیین کلاسی که نمونه داده‌شده متعلق به آن است، برای نمونه‌های دیده نشده بسیار بالا است، اما

---

[1] - Dependence Measures
[2] - Correlation
[3] - Consistency Measures
[4] - Classifier Error Rate Measures

هزینه‌های محاسباتی در آن‌ها نیز نسبتاً زیاد است.

در جدول زیر مقایسه‌ای بین انواع مختلف تابع ارزیابی، صرف‌نظر از نوع تابع تولیدکننده مورداستفاده، انجام‌شده است. پارامترهایی که برای مقایسه استفاده‌شده‌اند به‌صورت زیر می‌باشند:

۱. **عمومیت**[1]: اینکه بتوان زیرمجموعه انتخاب‌شده را برای طبقه‌بندی کننده‌های متفاوت به کار ببریم.

۲. **پیچیدگی زمانی**: زمان لازم برای پیدا کردن زیرمجموعه ویژگی جواب.

۳. **دقت**: دقت پیشگوئی با استفاده از زیرمجموعه انتخاب‌شده.

علامت "---" که در ستون آخر آمده است، به این معنی است که در مورد میزان دقت حاصل نمی‌توانیم مطلبی بگوییم. به‌جز خطای طبقه‌بندی کننده، دقت سایر توابع ارزیابی به مجموعه داده مورداستفاده و طبقه‌بندی کننده‌ای که بعد از انتخاب ویژگی برای طبقه‌بندی کلاس‌ها استفاده می‌شود، بستگی دارد.

جدول۲-۱ مقایسه توابع ارزیابی مختلف

نوع تابع ارزیابی	عمومیت	پیچیدگی زمانی	دقت
معیار فاصله	دارد	پائین	---
معیار اطلاعات	دارد	پائین	---
معیار وابستگی	دارد	پائین	---
معیار سازگاری	دارد	متوسط	---
خطای طبقه‌بندی کننده	ندارد	بالا	خیلی زیاد

۲-۵- دسته‌بندی و تشریح الگوریتم‌های مختلف انتخاب ویژگی

در این قسمت بر اساس تابع ارزیابی و تابع تولیدکننده، روش‌های مختلف انتخاب ویژگی را به چند دسته تقسیم‌بندی می‌کنیم که با ذکر منبع از توضیح روش‌هایی را که در این پایان‌نامه مورداستفاده قرار نگرفته‌اند گذر خواهد شد.

[1] - Generality

تابع ارزیابی مبتنی بر فاصله - تابع تولیدکننده مکاشفه‌ای[8]

تابع ارزیابی مبتنی بر فاصله - تابع تولیدکننده کامل[11]و[12]

تابع ارزیابی مبتنی بر اطلاعات - تابع تولیدکننده مکاشفه‌ای[15]

تابع ارزیابی مبتنی بر اطلاعات - تابع تولیدکننده کامل[16]

تابع ارزیابی مبتنی بر وابستگی - تابع تولیدکننده مکاشفه‌ای[18]و[19]

تابع ارزیابی مبتنی بر سازگاری - تابع تولیدکننده کامل [7]و20]

تابع ارزیابی مبتنی بر سازگاری - تابع تولیدکننده تصادفی[21]

تابع ارزیابی مبتنی بر خطای طبقه‌بندی کننده- تابع تولیدکننده مکاشفه‌ای

همان‌طور که قبلاً نیز اشاره کردیم، به مجموعه روش‌هایی که از تابع ارزیابی مبتنی بر نرخ خطای طبقه-بندی کننده استفاده می‌کنند، (بدون توجه به نوع تابع تولیدکننده استفاده‌شده) روش‌های wrapper می‌گویند. در این گروه روش‌های مشهور زیر را می‌توانیم ببینیم:

روش SFS (Sequential Forward Selection)

این روش، کارش را با یک مجموعه خالی شروع می‌کند، سپس در هر تکرار یک ویژگی با استفاده از تابع ارزیابی مورداستفاده، به مجموعه جواب اضافه می‌کند، این کار را تکرار می‌کند تا زمانی که تعداد ویژگی لازم انتخاب شود. مشکلی که این روش با آن روبروست، این است که ویژگی اضافه‌شده درصورتی‌که مناسب نباشد، از مجموعه جواب حذف نمی‌شود[23].

روش SBS (Sequential Backward Selection)

این روش برعکس SFS کارش را با مجموعه‌ای شامل تمام ویژگی‌ها شروع می‌کند و در هر بار تکرار

الگوریتم، ویژگی که به‌وسیله تابع ارزیابی انتخاب می‌شود، را از مجموعه موردنظر حذف می‌کند. این کار را تا زمانی ادامه می‌دهد که تعداد ویژگی‌ها برابر یک تعداد معینی شود. مانند روش قبل مشکل این روش این است که ویژگی حذف‌شده را دیگر به مجموعه اضافه نمی‌کند، حتی اگر مناسب باشد[23].

روش‌های دیگری که در این گروه وجود دارند، نسخه‌های متفاوتی از دو روش قبلی یا ترکیب آن‌ها هستند.

روش SBS-Slash

این روش بر پایه این مشاهده است که هنگامی‌که تعداد زیادی ویژگی وجود دارد، بعضی از طبقه‌بندی کننده‌ها (مانند ID3 یا C4.5) مکرراً تعداد زیادی از آن‌ها را استفاده نمی‌کنند. الگوریتم با یک مجموعه ویژگی کار خودش را شروع می‌کند (مانند SBS)، اما بعد از یک مرحله تمام ویژگی‌هایی را که در این مرحله یاد گرفته است و استفاده‌نشده‌اند، را حذف (Slashes) می‌کند[24].

روش PQSS ((p,q) Sequential Search)

در اینجا از بعضی از خواص بازگشت به عقب[1] استفاده می‌کنیم. عملکرد الگوریتم به این صورت است که در هر مرحله p ویژگی به مجموعه اضافه و q ویژگی از آن حذف می‌کند. حال اگر الگوریتم از مجموعه خالی شروع کرده باشد، بایستی اندازه p بزرگ‌تر از اندازه q باشد. ولی اگر از مجموعه تمام ویژگی‌ها شروع‌شده باشد، بایستی اندازه p کوچک‌تر از q باشد[25].

روش BDS (Bi-Directional Search)

مانند روش‌های قبل است با این تفاوت که جستجو را از دو طرف انجام می‌دهد[25].

---

[1] - Backtracking

روش Schemata Search

الگوریتم کارش را با مجموعه خالی و یا مجموعه تمام ویژگی‌ها شروع می‌کند و در هر تکرار، بهترین زیرمجموعه را با حذف یا اضافه تنها یک ویژگی به مجموعه ویژگی، پیدا می‌کند. برای اینکه هر زیرمجموعه را ارزیابی کند، از تعیین اعتبار (LOOCV)[1] استفاده می‌کند. در هر تکرار زیرمجموعه‌ای انتخاب می‌شود که کمترین خطای LOOCV را داشته باشد. کار به این صورت ادامه می‌یابد تا هیچ تغییر با تک ویژگی نتواند باعث بهتر شدن زیرمجموعه شود[26].

روش RC (Relevance in Context)

در اینجا این واقعیت تشریح شده است که بعضی از ویژگی‌ها فقط در قسمتی از فضای کار مربوط هستند. روش کار مشابه SBS است، اما با تغییرات عمده‌ای که باعث محلی شدن آن شده است(انتخاب ویژگی-های مرتبط بر اساس تصمیم‌گیری به‌وسیله نمونه‌ها)[27] .

روش Queiros and Gelsema

شبیه SFS است اما پیشنهاد می‌کند که در هر تکرار، هر ویژگی در با تنظیمات متفاوتی به‌وسیله اثرات متفاوت ناشی از ویژگی‌های قبلی ارزیابی شود[28]. دو نمونه از این تنظیمات به این صورت هستند:

همیشه فرض کنیم که ویژگی‌ها مستقل هستند (ویژگی‌های قبلی را در نظر نمی‌گیریم).

هیچ‌گاه فرض نمی‌کنیم که ویژگی‌ها مستقل هستند (ویژگی‌های قبلی را در نظر می‌گیریم).

در این روش و تعدادی از روش‌های قبلی در این گروه از نرخ خطای بیز به‌عنوان خطای طبقه‌بندی کننده استفاده می‌کنیم.

---

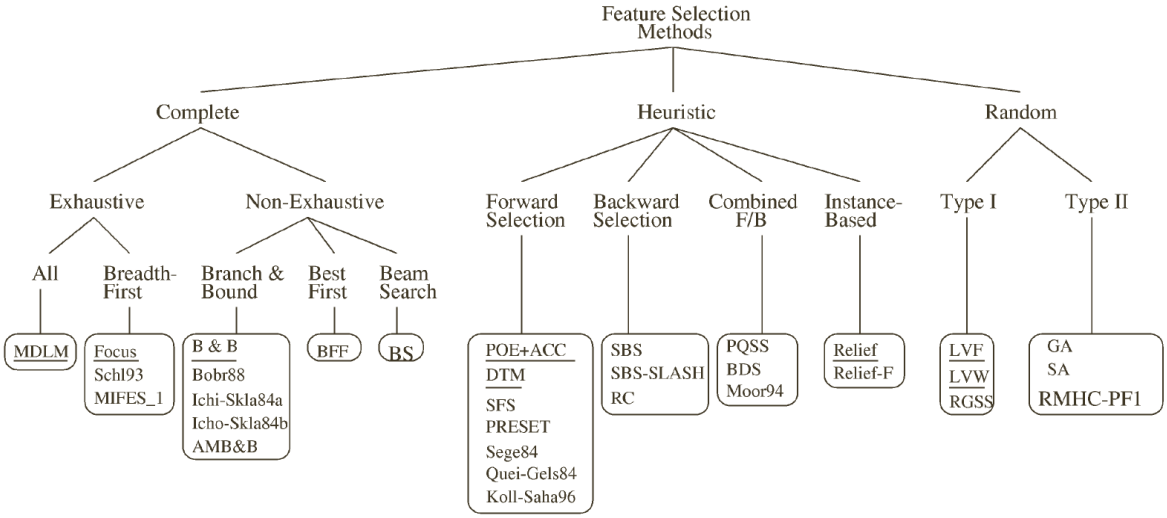
[1] Leave-One-Out Cross Validation

تابع ارزیابی مبتنی بر خطای طبقه‌بندی کننده - تابع تولیدکننده کامل [29,30,31]و [32]

تابع ارزیابی مبتنی بر خطای طبقه‌بندی کننده - تابع تولیدکننده تصادفی [33]

۲-۶- جمع‌بندی روش‌های انتخاب ویژگی

برای اینکه یک جمع‌بندی از کلیه روش‌های انتخاب ویژگی داشته باشیم، نمودار آن‌ها را برحسب سه نوع تابع تولیدکننده در شکل ۲-۴ نشان داده‌ایم[6].



شکل۲-۴ روش‌های مختلف انتخاب ویژگی

طبقه‌بندی

و اما روش کلاسه‌بندی که در الگوریتم پیشنهادی مورداستفاده قرارگرفته و در زیر به توضیح آن خواهیم پرداخت

۲-۷- ماشین بردار پشتیبان[1] :

ماشین بردار پشتیبان یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند.

این روش ازجمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی ازجمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کاری دسته‌بندی کننده SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. حل معادله پیدا کردن خط بهینه برای داده‌ها به‌وسیله روش‌های QP که روش‌های شناخته‌شده‌ای در حل مسائل محدودیت هستند صورت می‌گیرد. قبل از تقسیمِ خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به‌وسیله تابعِ phi به فضای با ابعاد خیلی بالاتر می‌بریم. برای اینکه بتوانیم مسئله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم از قضیه دوگانی لاگرانژ برای تبدیلِ مسئله مینیمم‌سازی موردنظر به فرم دوگانی آن‌که در آن به‌جای تابع پیچیده phi که ما را به فضایی با ابعاد بالا می‌برد، تابعِ ساده‌تری به نامِ تابع هسته که ضرب برداری تابع phi است ظاهر می‌شود استفاده می‌کنیم. از توابع هسته مختلفی ازجمله هسته‌های نمایی، چندجمله‌ای و سیگموید می‌توان استفاده نمود.

کاربردهای SVM

الگوریتم SVM، جز الگوریتم‌های تشخیص الگو دسته‌بندی می‌شود. از الگوریتم SVM، در هرجایی که نیاز به تشخیص الگو یا دسته‌بندی اشیاء در کلاس‌های خاص باشد می‌توان استفاده کرد. در ادامه به کاربردهای این الگوریتم به‌صورت موردی اشاره می‌شود:

[1] Support vector machines - SVMs

سیستم آنالیز ریسک، کنترل هواپیما بدون خلبان، ردیابی انحراف هواپیما، شبیه‌سازی مسیر، سیستم راهنمایی خودکار اتومبیل، سیستمهای بازرسی کیفیت، آنالیز کیفیت جوشکاری، پیش‌بینی کیفیت، آنالیز کیفیت کامپیوتر، آنالیز عملیات آسیاب، آنالیز طراحی محصول شیمیایی، آنالیز نگهداری ماشین، پیشنهاد پروژه، مدیریت و برنامه‌ریزی، کنترل سیستم فرایند شیمیایی و دینامیکی، طراحی اعضای مصنوعی، بهینه‌سازی زمان پیوند اعضا، کاهش هزینه بیمارستان، بهبود کیفیت بیمارستان، آزمایش اتاق اورژانس، اکتشاف روغن و گاز، کنترل مسیر در دستگاه‌های خودکار، ربات، جراثقال، سیستمهای بصری، تشخیص صدا، اختصار سخن، دسته‌بندی صوتی، آنالیز بازار، سیستمهای مشاوره‌ای محاسبه هزینه موجودی، اختصار اطلاعات و تصاویر، خدمات اطلاعاتی خودکار، مترجم لحظه‌ای زبان، سیستمهای پردازش وجه مشتری، سیستمهای تشخیص ترمز کامیون، زمان‌بندی وسیله نقلیه، سیستمهای مسیریابی، دسته‌بندی نمودارهای مشتری/بازار، تشخیص دارو، بازبینی امضا، تخمین ریسک وام، شناسایی طیفی، ارزیابی سرمایه، دسته‌بندی انواع سلول‌ها، میکروب‌ها و نمونه‌ها، پیش‌بینی فروش‌های آینده، پیش‌بینی نیازهای محصول، پیش‌بینی وضعیت بازار، پیش‌بینی شاخص‌های اقتصادی، پیش‌بینی ملزومات انرژی، پیش‌بینی واکنش‌های دارویی، پیش‌بینی بازتاب محصولات شیمیایی، پیش‌بینی هوا، پیش‌بینی محصول، پیش‌بینی ریسک محیطی، پیش‌بینی جداول داوری، مدل کردن کنترل فرآیند، آنالیز فعالیت گارانتی، بازرسی اسناد، تشخیص هدف، تشخیص چهره، انواع جدید سنسورها، دستگاه کاشف زیردریایی به‌وسیله امواج صوتی، رادار، پردازش سیگنال‌های تصویری شامل مقایسه اطلاعات، پیگیری هدف، هدایت جنگ‌افزارها، تعیین قیمت وضعیت فعلی، جلوگیری از پارازیت، شناسایی تصویر /سیگنال، چیدمان یک مدار کامل، بینایی ماشین، مدل کردن غیرخطی، ترکیب صدا، کنترل فرآیند ساخت، آنالیز مالی، پیش‌بینی فرآیندهای تولید، ارزیابی به‌کارگیری یک سیاست، بهینه‌سازی محصول، تشخیص ماشین و فرآیند، مدل کردن کنترل دستگاه‌ها، مدل کردن ساختارهای شیمیایی، مدل کردن سیستمهای دینامیکی، مدل کردن سیگنال تراکم، مدل کردن قالب‌سازی پلاستیکی، مدیریت قراردادهای سهام، مدیریت وجوه بیمه، مدیریت سهام،

تصویب چک‌بانکی، اکتشاف تقلب در کارت اعتباری، ثبت نسیه، بازبینی امضا از چک‌ها، پیش‌بینی ارزش نسیه، مدیریت ریسک رهن، تشخیص حروف و اعداء تشخیص بیماری و.....

خلاصه استفاده عملی از SVM

ماتریس الگو را آماده می‌کنیم. تابع هسته را برای استفاده انتخاب می‌کنیم. پارامتر تابع هسته و مقدار C را انتخاب می‌کنیم. برای محاسبه‌ی مقادیرα_i الگوریتم آمورشی را با استفاده از حل‌کننده‌های QP اجرا می‌کنیم. داده‌های جدید با استفاده از مقادیرα_i و بردارهای پشتیبان می‌توانند دسته‌بندی شوند.

مزایا و معایب SVM

آموزش نسبتاً ساده است برخلاف شبکه‌های عصبی در ماکزیمم‌های محلی گیر نمی‌افتد. برای داده‌های با ابعاد بالا تقریباً خوب جواب می‌دهد. مصالحه بین پیچیدگی دسته‌بندی کننده و میزان خطا به‌طور واضح کنترل می‌شود. به یک تابع هسته خوب و انتخاب پارامتر C نیاز دارد.

ماشین بردار پشتیبان خطی

ما مجموعه داده آموزش $\mathcal{D}$ شامل n عضو را در اختیارداریم که به‌صورت زیر تعریف می‌شود :

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n} \qquad (۱-۲)$$

جایی که مقدار y برابر ۱ یا -۱ و هر $\mathbf{x}_i$ یک بردار حقیقی p بعدی است. هدف پیدا کردن ابر صفحه جداکننده با بیشترین فاصله از نقاط حاشیه‌ای است که نقاط با $y_i = 1$ را از نقاط با $y_i = -1$ جدا کند. هر ابر صفحه می‌تواند به‌صورت مجموعه‌ای از نقاط $\mathbf{x}$که شرط زیر را ارضا می‌کند نوشت:

$\mathbf{w} \cdot \mathbf{x} - b = 0,$ که . علامت ضرب است و $\mathbf{w}$ بردار نرمال است، که به ابر صفحه عمود است. ما می-خواهیم $\mathbf{w}$و $\mathbf{b}$را طوری انتخاب کنیم که بیشترین فاصله بین ابر صفحه‌های موازی که داده‌ها را از هم جدا می‌کنند، ایجاد شود. این ابر صفحه‌ها با استفاده از رابطه زیر توصیف می‌شوند .

$$\mathbf{w} \cdot \mathbf{x} - b = 1 \tag{۲-۲}$$

و

$$\mathbf{w} \cdot \mathbf{x} - b = -1. \tag{۳-۲}$$

اگر داده‌های آموزشی جدایی‌پذیر خطی باشند، ما می‌توانیم دو ابر صفحه در حاشیه نقاط به‌طوری‌که هیچ نقطه مشترکی نداشته باشند، در نظر بگیریم و سپس سعی کنیم، فاصله آن‌ها را، ماکزیمم کنیم. با استفاده از هندسه، فاصله این دو صفحه $\frac{2}{\|\mathbf{w}\|}$است. بنابراین ما باید $\|\mathbf{w}\|$را مینیمم کنیم. برای اینکه از ورود نقاط به حاشیه جلوگیری کنیم، شرایط زیر را اضافه می‌کنیم: برای هر i

$$\text{of the first class } \mathbf{w} \cdot \mathbf{x}_i - b \geq 1 \qquad \text{for } \mathbf{x}_i \tag{۴-۲}$$

یا

$$\text{of the second class } \mathbf{w} \cdot \mathbf{x}_i - b \leq -1 \qquad \text{for } \mathbf{x}_i \tag{۵-۲}$$

این می‌تواند به‌صورت زیر نوشته شود:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1, \quad \text{for all } 1 \leq i \leq n. \tag{۶-۲}$$

با کنار هم قرار دادن این دو یک مسئله بهینه‌سازی به دست می‌آید:

$$\text{: Minimize (in } \mathbf{w}, b\text{)}$$

$$\|\mathbf{w}\| \tag{۷-۲}$$

$$\text{subject to (for any } i = 1, \ldots, n\text{)}$$

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1. \tag{۸-۲}$$

فرم اولیه

مسئله بهینه‌سازی مشاهده‌شده در قسمت قبل، مسئله سختی برای حل کردن است، زیرابه $\|\mathbf{w}\|$وابسته

است. خوشبختانه می‌توانیم، بدون تغییر در مسئله $\|\mathbf{w}\|$ را با $\frac{1}{2}\|\mathbf{w}\|^2$جانشین کنیم (عبارت ½ برای آسودگی در محاسبات ریاضی آورده شده). این‌یک مسئله بهینه‌سازی(OP) برنامه‌ریزی غیرخطی (QP) است. به‌طور واضح‌تر:

$$\textit{Minimize (in } \mathbf{w}, b\textit{)c}$$

$$\frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to (for any } i = 1, \ldots, n) \qquad (۹-۲)$$

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1. \qquad (۱۰-۲)$$

می‌توان عبارت قبل را با استفاده از ضرایب نا منفی لاگرانژ به‌صورت زیر نوشت که در آن ضرایب لاگرانژ هستند :α_i

$$\min_{\mathbf{w},b,\boldsymbol{\alpha}} \{\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i[y_i(\mathbf{w} \cdot \mathbf{x_i} - b) - 1]\} \qquad (۱۱-۲)$$

اما فرمول فوق اشتباه است. فرض کنید ما بتوانیم خانواده‌ای از ابر صفحات که نقاط را تقسیم می‌کنند پیدا کنیم. پس $y_i(\mathbf{w} \cdot \mathbf{x_i} - b) - 1 \geq 0$ بنابراین ما می‌توانیم مینیمم را با فرستادن همه α_iبه $+\infty$پیدا کنیم. بااین‌حال شرط پیش می‌تواند به‌صورت پایین بیان شود:

$$\min_{\mathbf{w},b} \max_{\boldsymbol{\alpha}} \{\frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i[y_i(\mathbf{w} \cdot \mathbf{x_i} - b) - 1]\} \qquad (۱۲-۲)$$

ما به دنبال نقاط saddle می‌گردیم. حالا می‌توان این مسئله را به کمک برنامه‌ریزی غیرخطی استاندارد حل کرد. جواب می‌تواند به‌صورت ترکیب خطی از بردارهای آموزشی بیان شود:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i} \qquad (۱۳-۲)$$

تنها چند α_iبزرگ‌تر از صفر خواهد بود. $\mathbf{x_i}$ متناظر، دقیقاً همان بردار پشتیبان خواهد بود و شرط را

ارضا خواهد کرد.

از این می‌توان نتیجه گرفت که بردارهای پشتیبان شرط زیر را نیز ارضا می‌کنند $y_i \ (w.x_i \ - b) =$ 1 که اجازه می‌دهد مقدار b تعریف شود. در عمل الگوریتم مقاوم‌تر خواهد بود اگر از تمام N_{SV} بردار پشتیبان میانگین گرفته شود:

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (\mathbf{w} \cdot \mathbf{x_i} - y_i) \qquad (۱۴-۲)$$

فرم دوگان

استفاده از این واقعیت که $\|\mathbf{w}\|^2 = w \cdot w$ و جانشینی $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i}$ می‌توان نشان داد که دوگان SVM به مسئله بهینه‌سازی زیر ساده می‌شود:

Maximize (in α_i*)*

$$\tilde{L}(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

subject to (for any $i = 1, \ldots, n$) (۱۵-۲)

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \qquad (۱۶-۲)$$

در اینجا هسته به‌صورت $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$ تعریف می‌شود. عبارت α تشکیل یک دوگان برای بردار وزن‌های مجموعه آموزشی می‌دهد:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i. \qquad (۱۷-۲)$$

ماشین بردار پشتیبان چند کلاسی

SVM اساساً یک جداکننده دودویی است. در بخش قبلی پایه‌های تئوری ماشین‌های بردار پشتیبان برای دسته‌بندی دو کلاس تشریح شد. یک تشخیص الگوی چند کلاسی می‌تواند به‌وسیله‌ی ترکیب ماشین‌های

بردار پشتیبان دو کلاسی حاصل شود. به‌طورمعمول دو دید برای این هدف وجود دارد. یکی از آن‌ها استراتژی "یک در مقابل همه " برای دسته‌بندی هر جفت کلاس و کلاس‌های باقی‌مانده است. دیگر استراتژی "یک در مقابل یک" برای دسته‌بندی هر جفت است. در شرایطی که دسته‌بندی اول به دسته‌بندی مبهم منجر می‌شود. برای مسائل چند کلاسی، رهیافت کلی کاهش مسئله‌ی چند کلاسی به چندین مسئله دودویی است. هریک از مسائل با یک جداکننده دودویی حل می‌شود. سپس خروجی جداکننده‌های دودویی SVM باهم ترکیب‌شده و به‌این‌ترتیب مسئله چند کلاس حل می‌شود.

ماشین‌های بردار پشتیبان غیرخطی

ابر صفحه جداکننده بهینه اولین بار توسط Vapnik در سال ۱۹۶۳ ارائه شد که یک دسته کننده خطی بود. در سال ۱۹۹۲ ،[۳۶] راهی را برای ایجاد دسته‌بند غیرخطی، با استفاده قرار دادن هسته برای پیدا کردن ابر صفحه با بیشتر حاشیه، پیشنهاد دادند. الگوریتم نتیجه شده ظاهراً مشابه است، به‌جز آنکه تمام ضرب‌های نقطه‌ای با یک تابع هسته غیرخطی جایگزین شده‌اند. این اجازه می‌دهد، الگوریتم، برای ابر صفحه با بیشترین حاشیه در یک فضای ویژگی تغییر شکل داده، مناسب باشد .ممکن است، تغییر شکل غیرخطی باشد و فضای تغییریافته، دارای ابعاد بالاتری باشد. به‌هرحال دسته کننده، یک ابر صفحه در فضای ویژگی با ابعاد بالا است، که ممکن است در فضای ورودی نیز غیرخطی باشد.

اگر از هسته با تابع گوسیین استفاده شود، فضای ویژگی متناظر، یک فضای هیلبرت نامتناهی است. دسته کننده‌ی بیشترین حاشیه، خوش‌ترتیب است، بنابراین ابعاد نامتناهی، نتیجه را خراب نمی‌کند. هسته‌های متداول به‌صورت زیر هستند:

چندجمله‌ای (همگن)

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j})^d \qquad (۱۸-۲)$$

چندجمله‌ای (ناهمگن)

$$k(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^d \quad (۱۹-۲)$$

گوسیین

for ،Radial Basis Function: $k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2)$

$\gamma > 0$.Sometimes parametrized using $\gamma = 1/2\sigma^2$ (۲۰-۲)

تانژانت هذلولی

for some (not every) ،: $k(\mathbf{x_i}, \mathbf{x_j}) = \tanh(\kappa \mathbf{x_i} \cdot \mathbf{x_j} + c)$

$\kappa > 0$and $c < 0$ (۲۱-۲)

هسته با انتقال $\varphi(\mathbf{x_i})$با تساوی

$$k(\mathbf{x_i}, \mathbf{x_j}) = \varphi(\mathbf{x_i}) \cdot \varphi(\mathbf{x_j}) \quad (۲۲-۲)$$

در ارتباط است. همچنین مقدار wدر فضای انتقال‌یافته برابر

$$\mathbf{w} = \textstyle\sum_i \alpha_i y_i \varphi(\mathbf{x}_i). \quad (۲۳-۲)$$

است. ضرب نقطه‌ای با wمی‌تواند توسط هسته محاسبه شود یعنی

$$\mathbf{w} \cdot \varphi(\mathbf{x}) = \textstyle\sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) \quad (۲۴-۲)$$

.به‌هرحال در حالت عادی 'wوجود ندارد، به‌طوری‌که

$$\mathbf{w} \cdot \varphi(\mathbf{x}) = k(\mathbf{w}', \mathbf{x}). \quad (۲۵-۲)$$

معیار صحت

اولین و پرطرفدارترین معیار استفاده‌شده در مقاله‌های مختلف معیار Hit ratio می‌باشد.

این فرمول در بحث دودسته جواب مشابهی با معیار Accuracy تولید می‌کند که در این تحقیق به علت دودسته بودن از معیار Accuracy استفاده می‌کنیم.

جدول۲-۲ ، معیارهای متفاوت حاصل از دودسته

		Condition as determined by *Gold* standard		
		True	**False**	
Test outcome	**Positive**	True positive	**False positive**	→ Positive predictive value or Precision
	Negative	**False negative**	True negative	→ Negative predictive value
		↓ Sensitivity or recall	↓ Specificity (or its complement, Fall-Out)	Accuracy

توضیحات مربوط به فرمول صحت در جدول ۲-۲ آمده است. به‌طور مثال True Positive و یا به‌صورت اختصار TP به معنای تعداد حالاتی است که خروجی دسته ۱ باشد و دسته‌بند نیز مقدار ۱ را به‌عنوان خروجی تخمین زده باشد.

$$Accuracy = ([1]Tp + [2]Tn)/(Tp + Tn + [3]Fp + [4]Fn) \qquad (۲۶-۲)$$

[1] True Positive
[2] True Negative
[3] False Positive
[4] False Negative

فصل سه

شناسایی و بررسی نشانگرها

۳- مقدمه

ساختار یک نشانگر بدین‌صورت است که بر اساس اطلاعاتی تعریفی و فرمول‌های مشخص یک نمودار که عمدتاً از جنس قیمت است ساخته می‌شود که در فضای تحلیل گری به دلیل نشانه‌هایی که برای تحلیل‌گر صادر می‌کند معروف به نشانگر یا اندیکاتور شده است. همان‌طور که در قیمت یک نقطه نمی-تواند بار مفهومی خاصی به همراه داشته باشد و ترکیب نقاط است که می‌تواند ایده بدهد در نشانگرها نیز مجموعه نقاط دوره‌های مختلف لازم است تا بار معنایی خاصی را تبیین کند.

نشانگرها بر اساس خلاقیت انسان‌ها و همچنین انتظاراتی که از ابزار دارند می‌تواند ساده یا پیشرفته باشد. به‌طورکلی نشانگرها برای اهداف متفاوتی طراحی می‌شوند و کاربردهای متفاوتی دارند، برخی ممکن است تائید یک حرکت باشند و برخی سیگنالی برای ورود به بازار و برخی دیگر نیز نشانه‌ای مبنی بر برگشت یک‌روند را تولید کنند. به‌طورکلی در تصمیم‌گیری و تشخیص روند نقش بسیاری دارند.

۳-۱- تعریف نشانگر

نشانگرها توابع ریاضی هستند که بر اساس فرمول‌های خاص در جهت تحلیل بازار به‌وسیله ابزارهای گرافیکی ترسیم و مورداستفاده قرار می‌گیرند. داده‌های ورودی در نشانگرها بر اساس تعدادی از بازه‌های قبلی تغییرات قیمت وارد می‌شود. در هر بازه زمانی اطلاعات قیمت آغاز(Open)، پایان(Close)، سقف(High)، کف(Low) و یا تعدادی از مجموع این اطلاعات مورداستفاده قرار می‌گیرد.

شکل ۳-۱نحوه‌ی محاسبه نشانگر

نشانگرها به لحاظ نوع حرکت و رفتارشان به دودسته کلی پیشرو[1] و تأخیری[2] تقسیم‌بندی می‌شوند: [۳۴]

نشانگرهای پیشرو معمولاً همراه با نوسانات قیمت هشدار و پیش‌بینی را اعلام می‌کنند و نشانگرهای تأخیری معمولاً با تأخیر پس از حرکت قیمت و جابجایی بازار هشدار و پیش‌بینی لازم را ارائه می‌دهند.

نشانگرهای پیشرو، اغلب موارد سیگنال‌های بیشتری نسبت به نشانگرهای تأخیری ایجاد می‌کنند و به همین جهت احتمال خطا در آن‌ها بیشتر است. نشانگرهای تأخیری بااینکه سیگنال‌های کمتری ایجاد می‌کنند اما به خاطر اینکه قیمت حرکت اصلی خود را در بازار انجام داده و سپس این نشانگرها سیگنال می‌دهند معمولاً تأخیر آن‌ها باعث از دست رفتن موقعیت مناسب ورود می‌شود. ایراد دیگر نشانگرهای تأخیری سیگنال‌های اشتباه آن‌ها در روندهای روبه‌جلو است. این دسته از نشانگرها معمولاً برای بازارها و بازه‌های زمانی مناسب هستند که بازار جهت و روند صعودی یا نزولی مشخص دارد.

۳-۱-۱- انواع نشانگرها

تقسیم‌بندی نشانگرها در مقالات و کتب بنا به نحوه استفاده و تخصص به چند خانواده اصلی بر اساس نوع

[1] Leading Indicator
[2] Lagging Indicator

نمایش یا مدل محاسباتی تقسیم‌بندی می‌شوند که مهم‌ترین این تقسیم‌بندی‌ها به‌صورت زیر است :

به‌طورکلی، نشانگرها را می‌توان به چهار گروه عمده طبقه‌بندی کرد[۱۷]:

- نشانگر شتاب
- نشانگر نوسانات
- نشانگر روند
- نشانگر مربوط به حجم

ده‌ها نشانگر، که بسیاری از آن‌ها با اندکی تغییرات در نشانگرهای معروف استفاده می‌کنند وجود دارد، اما اکثراً همان اطلاعات را منتقل می‌کنند. نشانگرهای انتخاب‌شده، نشانگر خوبی شناخته‌شده و یک نماینده خوبی برای تمام چهار گروه از نشانگرها می‌باشند.

در اغلب موارد، به‌جای استفاده از مقادیر نشانگر به‌طور مستقیم، به همان شکل مرسوم انجام می‌شود،

در حال حاضر با استفاده از روش تبدیل ارزش نشانگر به شکل تغییر در طی یک دوره از زمان انجام می‌شود. این اجازه می‌دهد تا ارتباط بین دو مقدار مختلف برای نشانگر (ثبت دونقطه مختلف در زمان)، به‌عنوان نرخ تغییرات نمایش داده شود (روند)[1] ، و آن را نیز هرگونه محاسبات ارزش نشانگر را در همان مقیاس قرار می‌دهد. به همین منظور روند حساسیت بیشتری نسبت به تغییرات ایجاد کرده، و از یک دوره نسبتاً کوتاه پنج روز از معاملات استفاده می‌کند.

برای تشخیص زودتر بازگشت روند، افزایش حساسیت نسبت به تغییر را فراهم می‌کند ، که این امر می-تواند منجر به سود بیشتر شود.

[1] Trend

۳-۱-۲- نشانگرهای مورداستفاده در الگوریتم

در این بخش به‌طور مفصل معرفی کامل و موارد استفاده معروف‌ترین نشانگرهایی که در اکثر نرم‌افزارها و به‌ویژه تحقیقات کاربرد دارند در این گزارش آمده است که در پیاده‌سازی هم مورداستفاده قرار گرفتند می‌پردازیم.

۱. *شاخص کانال کالا* ***CCI***[1]

شاخص کانال کالا (CCI) یک نشانگر چندمنظوره است که می‌تواند مورداستفاده برای شناسایی یک‌روند جدید و یا هشدار از شرایط شدید قرار گیرد. لم برت در اصل CCI را جهت تشخیص نوبت دوره‌ای در کالاها توسعه داد ، اما این نشانگر می‌تواند با موفقیت بر روی شاخص‌ها، صندوق‌ها، سهام و دیگر اوراق بهادار اعمال شود. به‌طورکلی، CCI سطح قیمت فعلی را نسبت به یک سطح قیمت متوسط در طی یک دوره معین از زمان اندازه‌گیری می‌کند. زمانی که قیمت به‌مراتب بالاتر از حد متوسط استCCI نسبتاً بالا است. همچنین زمانی که قیمت‌ها بسیار پایین‌تر از متوسط استCCI نسبتاً کم است. در این روش، CCI می‌تواند برای شناسایی سطوح خریدوفروش هیجانی مورداستفاده قرار می‌گیرد.

در مثال زیر که (CCI) برای یک دوره ۲۰ روزه محاسبه‌شده است. که تعداد دوره CCI برای محاسبه میانگین متحرک ساده و میانگین انحراف از نیز استفاده می‌شود.

CCI = (Typical Price - 20-period SMA of TP) / (.015 x Mean Deviation)

Typical Price (TP) = (High + Low + Close)/3Constant = .015

(۱-۳)

چهار مرحله برای محاسبه انحراف میانگین وجود دارد.

[1] Commodity Channel Index

اول، میانگین ۲۰ دوره قیمت معمول اخیر از هر دوره قیمت معمولی .

دوم، مقدار قدر مطلق این اعداد.

سوم، جمع مقادیر مطلق است.

چهارم، تقسیم‌بر تعداد کل دوره ۲۰.

با مقدار ثابت لم برت 0.015 به اطمینان می‌توان گفت که حدود ۷۰ تا ۸۰ درصد از ارزش CCIبین -100و 100+قرار می‌گیرند. این درصد همچنین به دوره قبل نیز بستگی دارد CCIکوتاه‌تر با دوره ۱۰ با درصد کمتری از بین ارزش‌های 100+ و 100- فرار خواهد بود در مقابل، CCIطولانی با دوره 40 با درصد بالاتری در بین ارزش‌های 100+و 100- خواهد بود.

۲. *نرخ تغییر ROC*[1]

نرخ تغییر، یک نوسان سنج مرکز محور است که نوسانات بالا و پایین صفر را نشان می‌دهد. همان‌طور که نام آن نشان می‌دهد، ROC درصد تغییر قیمت را در یک دوره مشخص اندازه می‌گیرد. به‌عنوان‌مثال، نرخ تغییر بیست‌روزه درصد تغییر قیمت را طی یک دوره ۲۰ روزه اندازه‌گیری می‌نماید. هرچه تفاوت بین قیمت جاری و قیمت ۲۰ روز گذشته بیشتر باشد، نرخ تغییر بیشتر خواهد بود. وقتی این نشانگر بالای صفر یا خط مرکزی است، درصد تغییر قیمت صعودی و مثبت است و بالعکس وقتی نشانگر زیر صفر باشد، درصد تغییر قیمت نزولی و منفی است. این نشانگر نیز به‌مانند MACD دارای محدوده بالا و پایین نیست که البته مشخصه تمام نوسان سنج‌های "مرکز محور" است و از طریق آن نشان دادن نقاط اشباع خریدوفروش، مشکل خواهد بود.

[1] Rate of Change

۳. شاخص قدرت نسبی RSI [1]

شاخص قدرت نسبی یا RSI در سال ۱۹۷۸ معرفی‌شده و از محبوبیت زیادی در میان معامله گران برخوردار است. RSI از نوع نوسانگرهایی است که سرعت و شتاب تغییرات قیمت را اندازه‌گیری می‌کند و همانند بیشتر نوسانگرها به بخش پایینی نمودار قیمت اضافه می‌شود. RSI بین ۰ تا ۱۰۰ نوسان می‌کند که دو سطح مهم آن یعنی ۳۰ و ۷۰ نشان‌دهنده سطوح "بیش فروش" (فروش افراطی) و "بیش خرید" (خرید افراطی) است

مفهوم "بیش خرید :"

زمانی که یک سهم یا نماد به‌صورت افراطی و غیرعادی خریداری می‌شود، نشانگر RSI آن به سمت سطوح بالای ۷۰ متمایل می‌شود(رسیدن به سقف). معامله‌گران در این شرایط به‌اصطلاح می‌گویند سهم "بیش خرید" شده است. در این زمان احتمال کاهش قیمت‌ها وجود دارد. هر چه شاخص RSI مدت بیشتری در سطح بالای ۷۰ باقی بماند احتمال کاهش قیمت نیز بیشتر می‌شود.

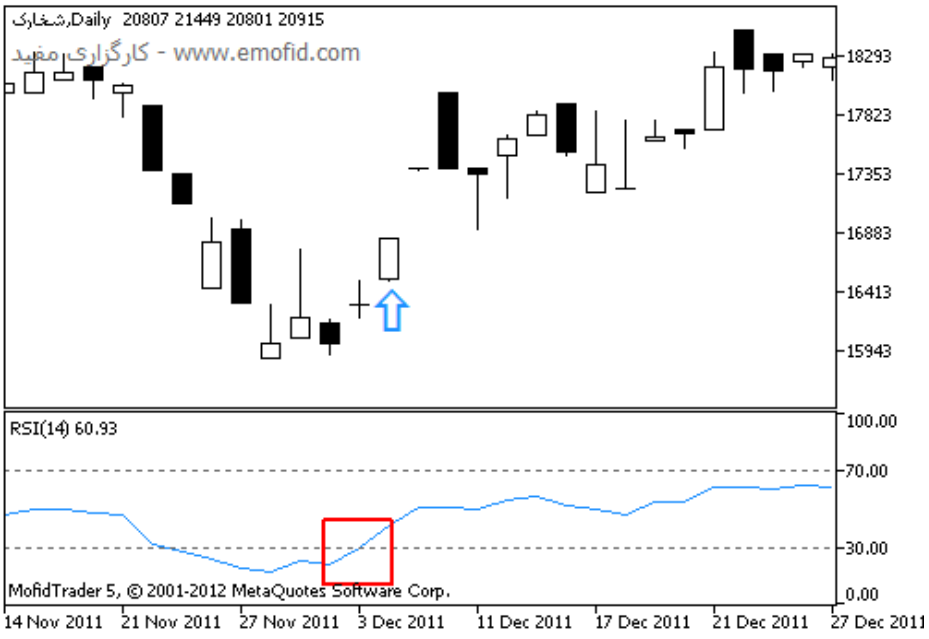
مفهوم "بیش فروش :"

زمانی که یک سهم یا نماد به‌صورت افراطی و غیرعادی فروخته می‌شود، شاخص RSI آن به سمت سطوح زیر ۳۰ متمایل می‌شود(رسیدن به کف). معامله گران در این شرایط به‌اصطلاح می‌گویند سهم "بیش فروش" شده است. در این زمان احتمال افزایش قیمت‌ها وجود دارد. هر چه نشانگر RSI مدت بیشتری در سطح زیر ۳۰ باقی بماند احتمال افزایش قیمت نیز بیشتر می‌شود.

[1] Relative Strength Index

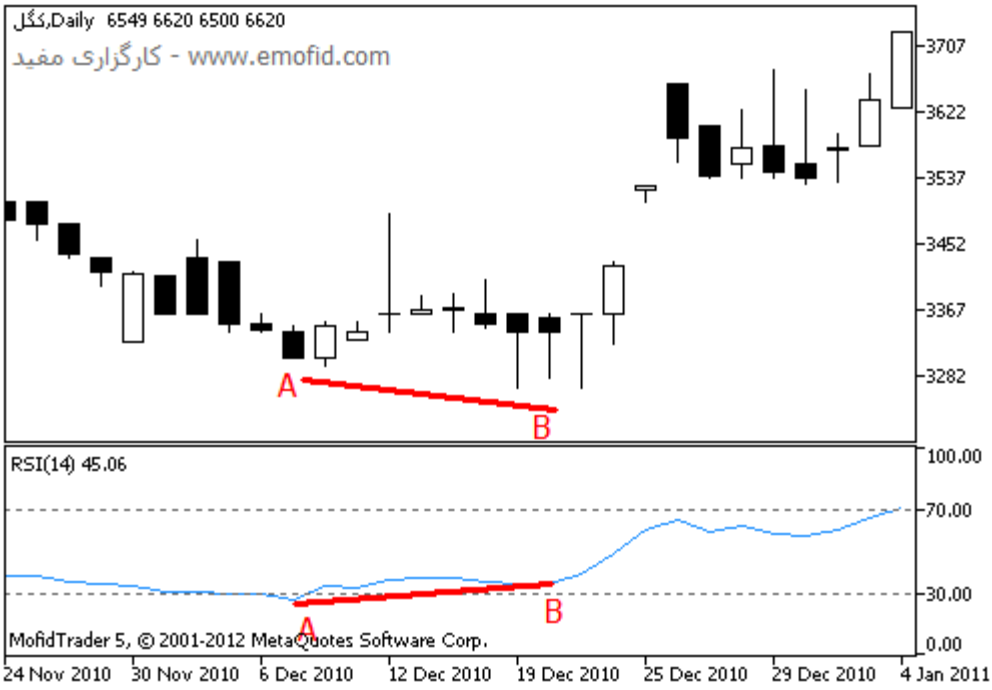
سیگنال‌های خرید در RSI

RSI برای سیگنال خرید ۱ :تقاطع صعودی RSI و سطح ۳۰ درصد. یعنی زمانی که RSI از زیر سطح ۳۰ درصد به سمت بالای این سطح حرکت می‌کند.



شکل ۳-۲سیگنال خرید rsi

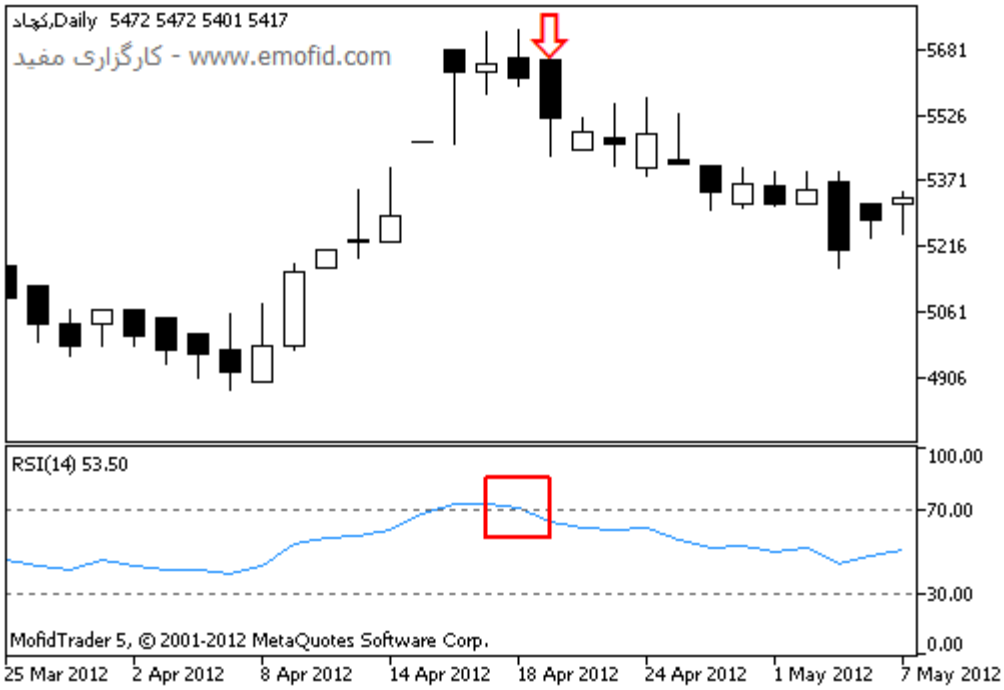
RSI برای سیگنال خرید ۲ :همگرایی میان نمودار قیمت و .RSI یعنی زمانی که در نمودار قیمت کف-های نزولی و در RSI کف‌های صعودی تشکیل می‌شود.



شکل۳-۳ سیگنال واگرایی در RSI

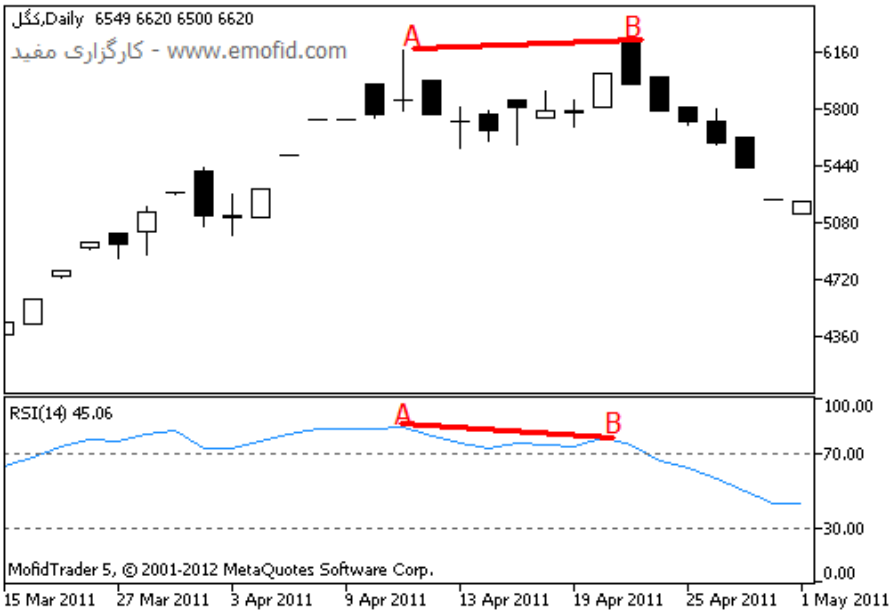
سیگنال‌های فروش در RSI

RSI برای سیگنال فروش ۱ :تقاطع نزولی RSI و سطح ۷۰ درصد. یعنی زمانی که RSI از بالای سطح ۷۰ درصد به سمت پایین این سطح حرکت می‌کند.



شکل ۳-۴سیگنال فروش در RSI

RSI برای سیگنال فروش ۲ : واگرایی میان نمودار قیمت و RSI . یعنی زمانی که در نمودار قیمت قله-های صعودی و در RSI قله‌های نزولی تشکیل می‌شود.



شکل ۳-۵ سیگنال واگرایی در RSI

و فرمول این نشانگر در زیر آمده است.

$$RSI = 100 - \frac{100}{1 + RS} \qquad (۲-۳)$$

$$RS = \text{Average Gain} / \text{Average Loss} \qquad (۳-۳)$$

۴. شاخص حجم توازنی OBV [1]

شاخص حجم توازنی (OBV) اولین بار توسط "جو گرنویل" عرضه شد که به‌منظور تائید روند جاری و صدور اخطار نسبت به تغییر جهت احتمالی روند به کار گرفته می‌شود. شیب و جهت OBV بسیار پراهمیت است و اعداد و مقادیر آن از اهمیت ثانویه برخوردارند. به‌طورکلی OBV می‌باید باقیمت همسو حرکت نماید، یعنی درروند قیمت صعودی، روند OBVنیز صعودی و بالعکس است. پس اگر میان OBV و قیمت واگرایی ملاحظه شود، به معنی سیگنال تغییر روند تفسیر می‌شود و احتمال تغییر جهت در آینده نزدیک زیاد است OBV. یک شاخص پیشرو (رهبر) تلقی می‌گردد زیرا تغییرات OBV معمولاً مقدم بر تغییرات قیمت است.

۵. نشانگر CMF [2]

اندیکاتور CMF توسط چکین به بازار معرفی‌شده است و به‌عنوان یک نوسان سنج با استفاده از خط تجمع و توزیع موردمحاسبه قرار می‌گیرد. خط تجمع (تقاضا) و توزیع (عرضه) بر این تعقل استوار است که میزان فشار خریدوفروش بر یک سهم را با توجه به قیمت پایانی و بالاترین و پایین‌ترین می‌توان در یک دوره مشخص شناسایی نمود. اگر قیمت بسته شدن یک سهم در نیمه بالای سطح نوسان قیمت روزانه

[1] On-Balance Volume
[2] Chaikin Money Flow

قرار گیرد، فشار خرید زیاد است و بالعکس اگر این قیمت در نیمه پایینی سطح نوسان قیمت روزانه قرار گیرد، فشار فروش زیاد می‌باشد .دوره زمانی را آقای چیکن بیست‌ویک روز توصیه می‌کند که درواقع فشار خریدوفروش یک ماه تقویمی است .حال اگر فشار فروش (توزیع)[1] زیاد باشد CMF کمتر از صفر است که این مسئله نشانگر نزولی بودن قیمت سهم است و فشار خرید و تجمع زمانی است که CMF بیشتر از صفر باشد که نشان از صعودی بودن قیمت دارد.

Money Flow Multiplier = [(Close - Low) - (High - Close)] /(High - Low) (۴-۳)

Money Flow Volume = Money Flow Multiplier x Volume for the Period

20-period CMF = 20-period Sum of Money Flow Volume / 20 period Sum of Volume

۶. *باندهای بولینگر BB*[2]

این نشانگر توسط جان بولینگر مطرح شد. در این نشانگر دو نوار (باند) در دو طرف یک میانگین متحرک نمایی گذاشته می‌شود. با این تفاوت که باندهای بولینگر با انحراف معیاری در بالا و پایین میانگین متحرکی که معمولاً ۲۰ دوره‌ای است، قرار می‌گیرند.

منظور از انحراف معیار یک مفهوم آماری است که بیان‌کننده چگونگی پراکندگی قیمت‌ها حول متوسط قیمت‌ها یا میانگین متحرک است. به‌عنوان یک قانون، زمانی که قیمت‌ها به باندهای بالایی برسند تمایل به بسط در جهت صعودی و زمانی که باند پایینی را لمس کنند، تمایل به بسط در جهت نزولی دارند.

یکی از ساده‌ترین کاربردهای باندهای بولینگر این است که باندهای بالا و پایین به‌عنوان هدف قیمتی در نظر گرفته شوند. به‌بیان‌دیگر اگر قیمت‌ها باند پایین را ترک کنند و میانگین متحرک ۲۰ روزه را رو به بالا قطع کنند مقدار باند بالایی را می‌توان به‌عنوان هدف قیمتی فرض کرد و به‌عکس.

---

[1] Distribution
[2] Bollinger Band

زمانی که باندهای بولینگر به صورت غیرمعمولی باز می‌شوند اکثراً نشانه این است که روند قبلی ممکن است به زودی تمام شود و زمانی که باندهای بولینگر به طور غیرمعمول به هم نزدیک شوند اکثرا نشانه این است که بازار آماده آغاز روندی جدید است.

شکل ۳-۶ بولینگر باند

درواقع تفسیر اصلی باندهای بولینگر این است که قیمت تمایل دارد در محدوده مابین دو باند بولینگر نوسان کند و به بیرون این محدوده نفوذ نکند؛ اما آنچه باندهای بولینگر را متمایز می‌کند این است که عرض باند با میزان نوسانات تغییر می‌کند. در دوره‌های نوسانات شدید قیمت، باندها با بیشتر شدن فاصله مابین خود به قیمت اجازه نوسان بیشتری می‌دهند و در دوره‌های راکد و بدون نوسان هم باندها به هم نزدیک شده و عرض کمتری را برای دربرداشتن نوسانات قیمت سهم فراهم می‌کنند.

باندهای بولینگر زمانی که با یکی از نوسانگرهای تشخیص خرید افراطی و یا فروش افراطی ادغام شوند، کاربرد بسیار خوبی پیدا می‌کنند. آقای بولینگر عدد ۲۰ را به‌عنوان بازه مناسب برای رسم میانگین متحرک معرفی می‌کند و نوع میانگین متحرک پیشنهادی ایشان ساده(Simple Moving Average)

است. از دیگر یافته‌های آقای بولینگر این است که میانگین متحرک‌های زیر ۱۰ روزه کارکرد خوبی برای رسم باندهای بولینگر ندارند.

Middle Band = 20-day simple moving average (SMA)

Upper Band = 20-day SMA + (20-day standard deviation of price x 2)

Lower Band = 20-day SMA - (20-day standard deviation of price x 2)

(۵-۳)

۷. شاخص جریان نقدینگی MFI[1]

نشانگر MFI که به معنی شاخص جریان (گردش) نقدینگی است MFI .درواقع اسیلاتوری است که با استفاده از قیمت و حجم معاملات ساخته می‌شود و معمولاً به‌عنوان RSIوزن داده‌شده با حجم شناخته می‌شود، این نشانگر شاخص جریان نقدینگی بین ۰ تا ۱۰۰ نوسان می‌کند و زمانی که تقاضا زیاد شده و قیمت رشد می‌کند به سمت بالا صعودی می‌شود و زمانی که قیمت افت کرده به عرضه زیاد می‌شود به سمت پائین نزول می‌کند، جالب است بدانید این نشانگر دقیقاً مثل نشانگر RSI است و تنها تفاوت محاسبه‌ای آن در استفاده از حجم در فرمول آن است. در مورد RSI زمانی که این به بالای مرز ۷۰ و یا زیر مرز ۳۰ قرار می‌گیرد به ترتیب اشباع خرید و اشباع فروش رخ می‌دهد که در مورد نشانگر MFI این موضوع به ترتیب اعداد ۸۰ و ۲۰ را شامل می‌شود و زمانی که مقدار MFI به بالای ۸۰ نفوذ می‌کند حالت اشباع خرید و زمانی که به زیر ۲۰ می‌رود اشباع فروش است.

$$typical\ price = \frac{high + low + close}{3} \quad (۶-۳)$$

$$money\ flow = typical\ price \times volume \quad (۷-۳)$$

[1] **M**oney **F**low **I**ndex

۸. نشانگر برد واقعی میانگین ATR[1]

نشانگر به‌عنوان ابزاری برای ارزیابی فراریت بازار معرفی‌شده است و فراریت به‌تنهایی از تلاش‌ها برای تعیین جهت جدا است. "برد واقعی میانگین" برخلاف اندیکاتور"برد واقعی (TR[2]) "فراریت فواصل و حرکات محدود را هم شامل می‌شود. این اندیکاتور ابزار خوبی است برای تخمین تمایل بازار برای حرکات قوی قیمت و جهش‌های ناگهانی که معمولاً بردهای وسیعی دارند.

"برد واقعی میانگین" در زمان‌بندی‌های روزانه و بالاتر در ۱۴ دورهٔ زمانی استفاده می‌شود که بازتاب‌دهندهٔ مقادیر فراریتی است که در ارتباط باقیمت ابزار معاملاتی هستند. مقادیر پایین "برد واقعی میانگین" با یک برد معاملاتی تطابق دارند درحالی‌که مقادیر بالا، ممکن است نمایانگر یک جهش ناگهانی صعودی یا نزولی باشد.

شکل ۳-۷ برد واقعی میانگین ATR

[1] Average True Range
[2] True Range

۹. نشانگر Aroon

نشانگرآرون توسط فردی به نام توشر در سال ۱۹۹۵ به وجود آمده است. این نشانگر به نسبت نشانگرهای دیگر جدیدتر می‌باشد. ویژگی اصلی آن این است که به‌طورکلی این را مشخص می‌کند که سهم روند دارد یا خیر؟ و درصورتی‌که روند داشته باشد، قدرت آن روند را تخمین می‌زند. که از دو قسمت جدا تشکیل‌شده است:

آرون بالا و پایین که در خط مرکزی با مقدار ۵۰ نوسان می‌کند و در محدوده بین ۰ تا ۱۰۰ قرار دارد. آرون بالابر پایه بیشترین قیمت می‌باشد و آرون پایین هم بر مبنای کمترین قیمت‌ها تنظیم‌شده است. زمانی که یک‌روند صعودی در حال تشکیل شدن می‌باشد آرون بالا ، بالای خط ۵۰ قرار می‌گیرد و آرون پایین در زیرخط ۵۰ قرار می‌گیرد. اما زمانی که روند سهم میل به نزولی شدن دارد، آرون پایین بالاتر از خط ۵۰ قرار می‌گیرد و آرون بالا در زیرخط ۵۰ قرار می‌گیرد.نحوه محاسبه:

Aroon-Up = ((25 - Days Since 25-day High)/25) x 100
Aroon-Down = ((25 - Days Since 25-day Low)/25) x 100

و معمولاً برای سهولت در مقایسه در کنار هم مشابه شکل۳-۸ نمایش داده می‌شوند.

شکل ۳- ۸ نمایش نشانگر *Aroon*

۱۰. نشانگر William's %R

نشانگر فنی دامنه درصد ویلیام (R٪) یک نشانگر فنی دینامیکی است که مشخص می کند که آیا بازار خرید اشباع/فروش اشباع است یا نه. R٪ ویلیام خیلی شبیه اسیلاتور استوکاستیک است. تنها تفاوت این است که R٪ یک مقیاس (اشل) معکوس دارد و اسیلاتور استوکاستیک دارای صاف کننده (رند کننده) داخلی است. برای نشان دادن نشانگر در این رسم معکوس، یک علامت منفی پیش از مقادیر دامنه درصد ویلیام (برای مثال ٪۳۰-) قرار داده‌شده است. باید از علامت منفی بهنگام انجام دادن تحلیل صرف‌نظر کرد.

نوسان کردن مقادیر نشانگر بین ۸۰ و ۱۰۰٪ نشان می‌دهد که بازار دچار حالت فروش اشباع‌شده است.

نوسان کردن مقادیر نشانگر بین ۰ و ۲۰٪ نشان می‌دهد که بازار دچار حالت خرید اشباع‌شده است.

۱۱. نشانگر Stochastic Oscillator

نشانگر Stochastic یا همان استوکاستیک در اواخر دهه ۱۹۵۰ معرفی‌شده و از نوع نوسانگرهایی است که سرعت و شتاب تغییرات قیمت را اندازه‌گیری می‌کند. که سه نسخه از این نشانگر وجود دارد یکی سریع(Fsto)[1] و یکی کند(Ssto)[2] و دیگری کامل که در این پایان‌نامه از دو نسخه سریع و کند استفاده‌شده. از نشانگر Stochastic می‌توان برای پیش‌بینی تغییر روند قیمت‌ها و یا مشخص کردن زمان بیش خرید و بیش فروش شدن یک سهم یا دارایی کمک گرفت. این نشانگر از یک منحنی اصلی و یک منحنی سیگنال تشکیل‌شده که مقادیر آن بین ۰ تا ۱۰۰ نوسان می‌کند. این نشانگر دو سطح مهم ۲۰ درصد و ۸۰ درصد دارد که نشان‌دهنده سطوح "بیش فروش" (فروش افراطی) و "بیش خرید" (خرید افراطی) است. نحوه محاسبه این نشانگر به‌صورت زیر است:

[1] Fast Stochastic Oscillator
[2] Slow Stochastic Oscillator

%K = (Current Close - Lowest Low)/(Highest High - Lowest Low) * 100
%D = 3-day SMA of %K

Lowest Low = lowest low for the look-back period
Highest High = highest high for the look-back period
%K is multiplied by 100 to move the decimal point two places

۱۲. نشانگر شاخص KDJ

شاخص KDJ درواقع یک‌شکل به‌دست‌آمده از اندیکاتور استوکاستیک است تنها با این تفاوت که با یک خط اضافی به نام خطJ شروع می‌شود. که خط J نشان‌دهنده واگرایی ارزش ٪ D از ارزش ٪K

۱۳. نشانگر EMA

میانگین متحرک نمایی با اعمال وزن بیشتری به قیمت‌های اخیر موجب کاهش تأخیر می‌شود. وزن اعمال‌شده به قیمت جدیدتر بستگی به تعداد دوره در میانگین متحرک دارد. سه‌گام برای رسیدن به محاسبه یک میانگین متحرک نمایی وجود دارد. اول، محاسبه میانگین متحرک ساده sma. دوم، محاسبه ضریب وزن. سوم، محاسبه میانگین متحرک نمایی. نحوه محاسبه به‌صورت زیر است:

SMA: 10 period sum / 10

Multiplier: (2 / (Time periods + 1)) = (2 / (10 + 1)) = 0.1818 (18.18%)
A: {Close - EMA(previous day)} x multiplier + EMA(previous day).

۱۴. نشانگر SMA

یک میانگین متحرک ساده از محاسبه میانگین قیمت تعدادی خاص از یک دوره تشکیل‌شده است. اکثر میانگین متحرک‌ها بر اساس قیمت بسته شدن محاسبه می‌شوند. برای مثال میانگین متحرک ساده برای پنج روز برابر است با مجموع پنج روز از قیمت بسته شدن تقسیم‌بر پنج. در شکل ۳-۹ این نشانگر نمایش داده‌شده است. نحوه محاسبه آن نیز به‌صورت مختصر به این نحو است:

Daily Closing Prices: 11,12,13,14,15,16,17

First day of 5-day SMA: (11 + 12 + 13 + 14 + 15) / 5 = 13

Second day of 5-day SMA: (12 + 13 + 14 + 15 + 16) / 5 = 14

Third day of 5-day SMA: (13 + 14 + 15 + 16 + 17) / 5 = 15

شکل ۳- ۹ نمایش نشانگر EMA,SMA

۱۵. نشانگر MACD

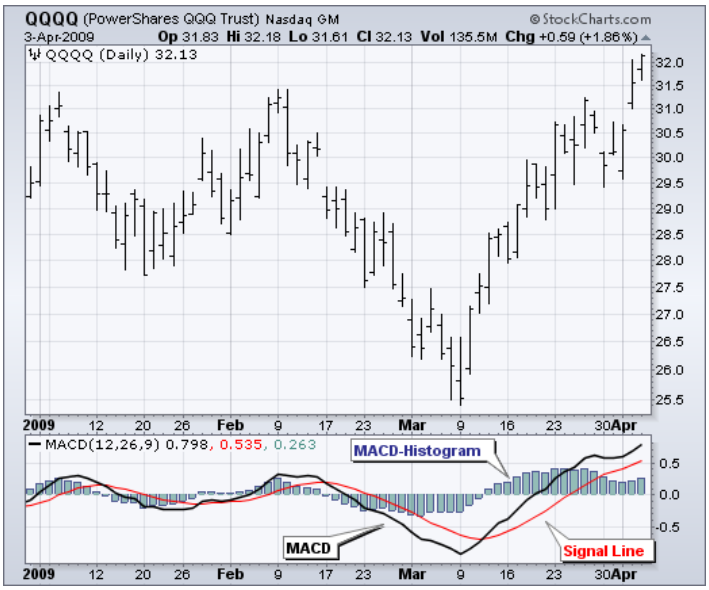
میانگین متحرک همگرا/واگرا یا MACD در اواخر دهه هفتاد میلادی معرفی‌شده و از آن زمان تاکنون جزو اندیکاتورهای محبوبِ معامله گران است .جزو اندیکاتورهایی است که شتاب و اندازه حرکت بازار را نشان می‌دهد مشابه (RSI) اما به خاطر طراحی ویژه آن می‌تواند برای تعیین روند بازار هم مورداستفاده قرار گیرد.

اندیکاتور MACD از ترکیب یک نمودار میله‌ای (هیستوگرام) و یک منحنی سیگنال تشکیل‌شده است. نمودار میله‌ای از تفاضل مقادیر دو میانگین متحرک سریع (۱۲دوره‌ای) و کند (۲۶دوره‌ای) به دست می‌آید. منحنی سیگنال میانگین متحرک سریع(۹ دوره‌ای) از تغییرات نمودار میله‌ای است. در شکل ۳-۱۰ این نشانگر نمایش داده‌شده است. نحوه محاسبه :

MACD Line: (12-day EMA - 26-day EMA)

Signal Line: 9-day EMA of MACD Line

MACD Histogram: MACD Line - Signal Line



شکل ۳- ۱۰ نمایش نشانگر MACD

۱۶. نشانگر ADX

اندیکاتور فنی شاخص حرکت جهت‌دار متوسط (ADX) به تشخیص روند و قدرت روند کمک می‌کند. این شاخص توسط آقای Welles Wilder در کتابش "New concepts in technical trading systems" شرح داده شده است. که چگونه به‌وسیله مقایسه اختلاف بین دو کف و دو سقف، قدرت روند و همچنین حرکات احتمالی قیمتی بیشتری را تخمین می‌زند.

۱۷. نشانگر Triple EMA (T3)

این شاخص فنی برای هموار کردن قیمت‌ها و دیگر داده‌ها استفاده می‌شود. که ترکیبی است از یک میانگین متحرک نمایی تنها و یک میانگین متحرک نمایی دو برابر و یک میانگین متحرک نمایی سه برابر شده.T3 نوسانات قیمت را هموار و فیلتر می‌کند، درنتیجه با تأخیر کمی آن را برای شناسایی روند آسان‌تر می‌کند. این‌یک ابزار مفید در شناسایی قوی، روند طولانی‌مدت است، اما ممکن است استفاده از آن محدود به، نوسانات کوتاه‌مدت بازار باشد.

۱۸. نشانگر شاخص قدرت واقعی TSI

شاخص قدرت واقعی (TSI) یک اسیلاتور شتاب بر اساس دو سطح از تغییرات قیمت است و باوجود اینکه چند مرحله برای محاسبه نیاز دارد، درواقع نشانگری بسیار ساده است. با تغییرات قیمت هموار، TSI فرازوفرود رفتار قیمت را با یک خط ثابت‌تر که فیلتر نویز می‌باشد ثبت می‌کند.

۳-۲- نتیجه‌گیری

با توجه به اهمیت و نقش نشانگرها و اینکه استفاده از آن‌ها چه کمکی در مشخص کردن روند بازار می-تواند داشته باشد، به همان نسبت نیز تعیین اینکه کدام‌یک از آن‌ها (چه تعدادی از این نشانگرها) اضافی یا تأثیر کمتری دارند شاید مؤثر باشد. با دخیل کردن تعداد زیادی از این نشانگرها جز پیچیدگی استدلال‌هایی که در پس این انتخاب قراراست صورت گیرد کمک دیگری در پیش‌بینی نخواهد کرد. مسلماً سهم‌های مختلف با نوسانات و سطوح مختلف قیمت و حجم معاملات ، هرکدام تحلیل و نشانگر مختص خود را می‌طلبد. چراکه هر نشانگر به‌طور مجزا یک دیدگاه و اطلاعات خاص خود را نسبت به دیگری می‌تواند در برداشته باشد. لذا از طرف دیگر محدود کردن کارگزار و یا سهام‌داران تنها به تعداد محدودی از نشانگرها نیز شاید منصفانه نباشد، به این خاطر که هر شخص تحلیل و استدلال و روش مختص خود را در استفاده و تنوع نشانگرها می‌تواند داشته باشد، پس شایسته‌تر هست که تنها به حذف آن تعداد از نشانگرها که در پیش‌بینی هیچ تأثیر و عملکردی ندارند قناعت کنیم.

فصل چهار

الگوریتم پیشنهادی

مقدمه

با در نظر گرفتن تأثیرات بنیادین در بازار بورس ایران از دیدگاه کارگزاران که شاید از ثبات پایداری برخوردار نباشد و اینکه هر تحلیل و پیش‌بینی نیازمند دانش بنیادین و همچنین فنی می‌باشد. و هر الگوریتمی که از جنبه‌ی فنی هدفی در راستای کمک به کارگزاران و تحلیلگران در جهت پیش‌بینی روزهای آتی بازار بورس داشته باشد بر پایه نشانگرها و محاسبات مبتنی بر قیمت خواهد بود. لذا هر الگوریتمی که با این هدف بتواند بستر مناسبی را برای سهامداران در جهت اخذ چنین انتخاب فراهم نماید. از استقبال و اهمیت برجسته‌ای برخوردار خواهد بود. الگوریتم پیشنهادی در این پایان‌نامه بعد از پیاده‌سازی نشانگرهای کاربردی با یک روش انتخاب ویژگی که در انتخاب نشانگرها مورداستفاده قرارگرفته، و در مرحله بعد با استفاده از یک روش طبقه‌بندی میزان صحت نشانگر انتخابی را تائید می‌کند. که درنهایت می‌تواند تعداد مؤثری از نشانگرها را برای تحلیل فنی پیشنهاد دهد. همان‌طور که تعیین نشانگرهای مفید در پیش‌بینی مؤثر هستند، تعیین نشانگرهایی که بی‌تأثیر هستند می‌تواند مفید باشد. چراکه در این مواقع دخیل کردن نشانگرهایی که نمی‌توانند کمکی در پیش‌بینی کنند موجب پیچیدگی و گمراه کردن کاربر خواهد شد.

۴-۱- اهمیت انتخاب ویژگی در پیش‌بینی

در راستای نشان دادن اهمیت اینکه انتخاب نشانگرهای مؤثر چه تأثیری می‌توانند در پیش‌بینی بازار بورس ایران داشته باشند. ابتدا به‌عنوان نمونه از ۳۰ نشانگری که در این پایان‌نامه به آن‌ها پرداخته‌شده است، اهمیت و نحوه کاربرد دو تا از آن‌ها را با ذکر مثال و شکل بر روی سهام سیمان شاهرود تشریح می‌کنم. لازم به ذکر است این نمودار قیمت مربوط به نمودار قیمت روزانه سهام شرکت سیمان شاهرود از تاریخ بهمن ۱۳۹۲ تا مردادماه ۱۳۹۴می‌باشد.

شکل۴-۱ قیمت روزانه سهام شاهرود و چند نشانگر

برای مثال با اعمال و محاسبه نشانگر میانگین متحرک با دوره ۱۴ روزه بر روی قیمت بسته شدن روزانه سهام سیمان شاهرود با نماد سرود(اعمال نشانگر بر روی قیمت بسته شدن بستگی به سلیقه و نظر کارگزار می‌تواند به شکل دیگری ظاهر شود) که بارنگ قرمز در شکل۴-۱ نمایش داده‌شده است. یکی از نشانگرهای مهم و مورداستفاده فراوان، در تحلیل تکنیکی است، که با حذف نوسانات قیمتی کمک می‌کند، تا سرمایه‌گذار بتواند تصویر بهتری از متوسط قیمت و روند قیمتی را ترسیم کند.

اولاً این نشانگر پایه محاسبه چندین نشانگر دیگر می‌باشد. و ازآنجایی‌که نسبت به نمودار قیمت تأخیر دارند. در گروه نشانگرهای تعقیب‌کننده روند قیمت قرار می‌گیرند.

وقتی قیمت‌ها دارای روند باشند این نشانگرها عملکرد خوبی دارند، و درصورتی‌که قیمت‌ها فاقد روند باشند میانگین متحرک گمراه‌کننده خواهد بود. حالا الگوریتمی که بتواند مشخص کند چه زمانی نشانگر مفید هست و چه زمان گمراه کننده، مطمئنا کمک شایانی به کاربر خواهد نمود.

همچنین در مثال دوم نشانگر بلینگر باند که با دوره ۲۰ روزه و میزان انحراف معیار۳ و بارنگ زرد

در شکل نشان داده‌شده است، اطلاعاتی را به سهام‌دار منتقل می‌کند که بتواند در پیش‌بینی روند قیمت مؤثر باشد.

با فرض اینکه این نشانگر می‌تواند در پیش‌بینی تأثیر گزار باشد، و الگوریتم ما به‌عنوان مؤثرها آن‌ها پیش نهاد کرده است، به کاربرد آن در این قسمت می‌پردازیم.

این نشانگر به کاربران این امکان را می‌دهد در طی یک مدت‌زمان از صعود یا نزول قیمت در آینده مطلع شوند و یا با ترکیب نشانگرهای دیگر و یا قیمت واقعی به پیش‌بینی آینده دست‌یافت، به این صورت که نزدیک شدن بیش‌ازحد باندها به یکدیگر نشان‌دهنده شروع یک‌روند جدید است و فاصله گرفتن باندهای بالا و پایین از همدیگر حاکی از اتمام روند قبل است.

همچنین برخورد نمودار قیمت‌ها به باندهای پایینی نشانی از بیش فروش شدن دارایی است و در این حالت احتمال افزایش قیمت‌ها بیشتر می‌شود و باندهای بالایی را می‌توان به‌عنوان سطوح مقاومت در نظر گرفت. برخورد قیمت‌ها به باندهای بالایی نشانی از بیش خرید شدن دارایی است و در این حالت احتمال کاهش قیمت‌ها بیشتر می‌شود. در این مرحله نیز به این مهم دست می‌یابیم که تأیید مفید واقع‌شدن این نشانگر از اهمیت بالایی برخوردار است.

۴-۲- الگوریتم پیشنهادی:

در اینجا الگوریتمی برای هر سهام خاص از بازار طراحی‌شده تا به‌واسطه آن مهم‌ترین و مؤثرترین نشانگرها کشف و مورداستفاده قرار گیرند. که درنهایت میزان صحت نشانگرهای منتخب توسط این الگوریتم با مقدار صحت(accuracy) هر سهام مشخص خواهد شد.

به این صورت که در ابتدا داده‌های سهام موردنظر شامل قیمت بالاترین[۱]، پایین‌ترین[۲]، قیمت بسته شدن[۱]

[1] high
[2] low

،باز شدن[2] و حجم معاملات[3] هرروز بازار از نرم‌افزار سهام ایران یا همان متاتریدر استخراج‌شده سپس داده-ها وارد متلب می‌شوند در فاز بعد بر روی‌داده‌ها خام فرمول‌هایی یا همان نشانگرها اعمال‌شده تا یک سری ویژگی مرسوم را که معامله‌گران بازار بورس برای ورود و خروج در بازار مورداستفاده قرار می‌دهند استخراج شود، ما به هرکدام از این نشانگرها به دید یک ویژگی نگاه کرده و به دنبال شناسایی بهترین آن‌ها هستیم.

هر نشانگر یک اطلاعات مجزا و متفاوتی نسبت به یک سهم دارد برای مثال برای اینکه کارگزار از روند قیمت سهم در بازار مطلع شود می‌تواند با اعمال نشانگرهایی مثلAroon و یا بلینگر و تحلیل آن‌ها به نتیجه مطمئن‌تری دست یابد.

با در نظر گرفتن اینکه ابتدا تمام نشانگرها مؤثر و دخیل هستند، کل نشانگرهای استخراج‌شده را به ماشین بردار پشتیبان داده و مقدار صحت پیش‌بینی فردا را با استفاده از روش 10-fold پیدا می‌کنیم(که ۱قسمت جهت آزمایش و مابقی جهت آموزش می‌باشد، عملیات آزمون و آموزش را توسط دسته‌بند کننده ماشین بردار پشتیبان انجام می‌دهیم و همین روال برای کل ۱۰ بخش تکرار می‌شود)

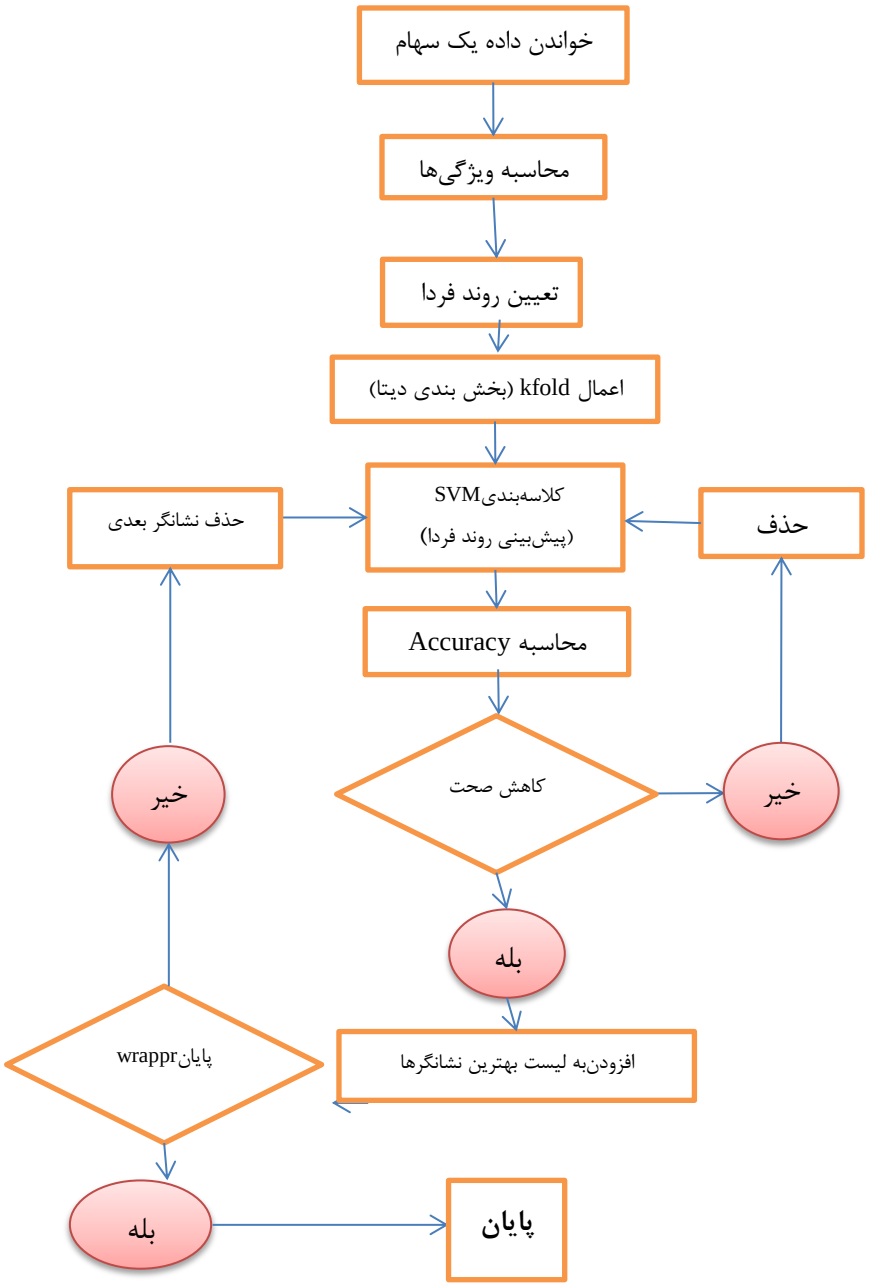
در مرحله بعد با استفاده از روش wrapper یک نشانگر را حذف و مجدداً مرحله آموزش و آزمون را اجرا می‌کنیم و درنهایت میزان صحت (accuracy) را برای نشانگرهای باقی‌مانده دریافت می‌کنیم. مقدار صحت کلی که در مرحله اول با کل نشانگرها به دست آورده بودیم را با مقدار صحت محاسبه‌شده پس از حذف نشانگر موردنظر، مقایسه می‌کنیم. اگر مقدار صحت نسبت به مرحله قبل کاهش‌یافته بود، به این معنی است که حذف نشانگر موردنظر باعث کاهش دقت پیش‌بینی‌شده و آن نشانگر تأثیر مثبتی در پیش-بینی داشته است. با این مقایسه می‌توان به میزان تأثیرگذاری و دخیل بودن تک‌تک نشانگرها در پیش-

---

[1] close
[2] open
[3] volume

بینی دست‌یافت. همچنین برای مقایسه دقیق‌تر می‌توان از معیارهای دیگر این دسته‌بند کننده(f-measure, false alarm, presision, …) که در خروجی قابل‌دسترس هستند استفاده نمود.

کلیت مراحل این الگوریتم را می‌توان در فلوچارت ۴-۲ مشاهده کرد.



شکل۴-۲ مراحل الگوریتم پیشنهادی

۴-۳- نتایج :

در این بخش ابتدا به مقایسه‌ای از خروجی‌های الگوریتم پیشنهادی و نشانگرهایی که توسط آن به‌عنوان بهترین‌ها و مؤثرترین‌ها انتخاب‌شده‌اند پرداخته‌شده است.

خروجی‌ها از سهام خاص و معتبر انتخاب‌شده. (نرم‌افزار مورداستفاده و نحوه‌ی گرفتن داده در بخش ضمایم آورده شده). برای نمونه داده‌های قیمت روزانه سهام شرکت سیمان شاهرود (که شامل حجم معاملات، قیمت باز شدن، قیمت بسته شدن، سقف قیمت و کف قیمت در روزاست) را برای مدت ۳۰ ماه به‌عنوان ورودی به الگوریتم دادیم تا ابتدا تمام نشانگرهای برآمده از قیمت این سهام محاسبه‌شده و صعود و نزولی بودن قیمت نسبت به‌روز قبل تعیین می‌شود و سپس بعد از نرمالایز کردن دیتا، خروجی به‌دست‌آمده را جهت بررسی میزان تأثیر هر نشانگر در پیش‌بینی روند با استفاده از کلاسه‌بندی توسط ماشین بردار پشتیبان به الگوریتم داده، و درنهایت دو خروجی از الگوریتم خواهیم داشت. یکی نشانگرهایی که در پیش‌بینی مؤثر بودند دوم میزان صحت پیش‌بینی برای سهام می‌باشد.(به این صورت که ابتدا با داشتن روند قیمت فردا با مقدار ۱ و ۱- سعی در پیش‌بینی صعودی یا نزولی بودن آن با استفاده از نشانگرها می‌کنیم.)

درنهایت خروجی سهام شرکت سیمان شاهرود،که نشانگرهای مؤثر و میزان صحت پیش‌بینی در آن مشخص‌شده به‌صورت زیر می‌باشد.

نشانگرهای انتخاب‌شده :

CCI, RSI, KDJ, SSTO, FSTO, Willr, TSI, EMA, MACD-Signal, ADX, OBV, CMF, Force, MFI, BB, Keltner, ATR, Change, SAR

۱۹ نشانگر بالا مؤثرترین نشانگرهای سهام شرکت سیمان شاهرود هستند با میزان صحت پیش‌بینی ۷۷%

. لازم به ذکر است که نشانگرهای پیشنهادی توسط الگوریتم با این فرض که تأثیری هرچند ناچیز در امر

پیش‌بینی خواهند داشت، پیشنهادشده‌اند. و به این معنی نیست که استفاده از آن‌ها ضروری هست. درواقع سرمایه‌گذاران را هنگام استفاده از این نشان‌گرها تأیید می‌کند.

سهام دیگری که به‌عنوان نمونه موردبررسی قرارگرفته و بعد از محاسبه نشانگرها و کلاس‌بندی خروجی با مقدار صحت ۸۱% را نشان می‌دهد مربوط به قیمت سهام نفت پارس می‌باشد، خروجی برای ۸۲۲ داده نفت پارس بهترین نشانگرهای انتخاب‌شده :

RSI, KDJ, SSTO, FSTO, Dn-Up, TSI, SMA, EMA, MACD, MACD-Signal, T3, OBV, CMF, MFI, BB, Keltner, ATR, Change, SAR

که ۱۹ نشانگر نهایتاً به‌عنوان مؤثرها برگزیده شدند. البته لازم به ذکر است محاسبه تعداد زیادی از نشانگرها به یکدیگر وابسته می‌باشد. یا ترکیب چند نشانگر می‌تواند خود سازنده نشانگری جدید باشد که این موضوع در فصل در بخش محاسبه نشانگرها مشهود است.

۴-۴- مقایسه با دیگر روش‌ها:

مقالات نوشته‌شده درزمینهٔ سهام ایران مربوط به پیش‌بینی قیمت می‌باشند و درصد خطای پیش‌بینی قیمت را به‌عنوان خروجی نمایش داده‌اند، به علت پیچیده بودن بازار سهام و پیچیده بودن پیش‌بینی قیمت سهام و زیاد بودن عوامل تأثیرگذار ما به دنبال صعود و یا نزول قیمت در فردای بازار هستیم. به علت نبود مقاله‌ای درزمینه‌ی پیش‌بینی روند فردای بازار سهام ایران جهت مقایسه با این کار روش‌های دیگری نیز پیاده شده و با داده‌ی یکسان با روش پیشنهادی مورد مقایسه. که در جدول ۴-۱ به نمایش گذاشته‌شده است.

نام روش	مقدار صحت	میزان انحراف از معیار	ویژگی‌های انتخاب‌شده
روش پیشنهادی	٪۸۸	۲,۱۹	RSI, KDJ, SSTO, FSTO, TSI, SMA, EMA, MACD, MACD-Signal, ADX, OBV, CMF, Force, MFI, BB, Keltner, ATR, Change, SAR
روش نزدیک‌ترین همسایه	٪۷۸	۳,۳۴	RSI, KDJ, FSTO WillR, SMA, Force, MFI, Keltner, Change

جدول۴-۱، مقایسه‌ی روش پیشنهادی با روش نزدیک‌ترین همسایه

۴-۵- نتیجه‌گیری:

با توجه به آزمایش‌های انجام‌شده روی سهام مختلف نشانگرهای زیر در بیشتر سهام تأثیرگذار بوده به همین جهت آن‌ها را به‌عنوان بهترین نشانگرهای بازار در حال حاضر شناسایی نمودیم که به شرح زیر می‌باشند.

RSI, KDJ, SSTO, FSTO, TSI, EMA, MACD-Signal, OBV, CMF, MFI, BB, Keltner, ATR, Change, SAR, darsad(cci), darsad(sma), darsad(ema), darsad(macd), darsad(adx)

همان‌طور که از نتایج مشخص است، این نشانگرها به همراه ماشین بردار پشتیبان توانسته به‌صورت خوبی فردای بازار را مورد پیش‌بینی قرار دهد، درنهایت نشانگرهای انتخاب‌شده می‌توانند توسط معامله‌گران مورداستفاده قرارگرفته و این پژوهش کمک می‌کند که از سردرگمی موجود در بازار سهام در جهت پیدا کردن بهترین نشانگرها کاسته شود. و همان‌طور که انتخاب نشانگرهای مؤثر می‌تواند از اهمیت بالایی برخوردار باشد به همان نسبت نیز مسلماً مشخص کردن نشانگرهایی که موجب گمراهی سهامداران می‌شوند اهمیت زیادی دارند.

۴-۶- پیشنهادها:

در کنار تمامی مراحل الگوریتم پیشنهادی و همچنین رعایت پیاده‌سازی نشانگرهای استفاده‌شده در الگوریتم می‌توان با گسترش دامنه تعداد نشانگرهای موجود در بازارهای جهانی به درجه‌ی بالاتری از صحت پیش‌بینی روند بازار دست‌یافت.

همچنین برای سهام خاصی که به تعداد متفاوت‌تری از نشانگرها (کمتر یا بیشتر) جهت تشخیص روند یا قیمت بازار نیاز دارند می‌توان از دیگر روش‌های انتخاب ویژگی و یا تابع هسته چندجمله‌ای و... بهره گرفت. و یا در صورت لزوم برای استفاده از همین روش می‌توان یک مقدار مشخص برای صحت تعیین کرد. به این صورت که اگر به تعداد نشانگر کمتری در خروجی نیاز داریم، هنگام مقایسه مقدار نزدیکی

صحت قبل با صحت جدید، از دقت کمتری استفاده کنیم.

۴-۷- ضمایم :

نرم‌افزار متاتریدر ۴ رسا که جهت دریافت داده‌های به‌روز سهام مختلف استفاده شد. که اسکریپتی برای دریافت خروجی اکسل از نرم‌افزار نوشته شد. یعنی قیمت روزانه هر سهام که شامل قیمت بالاترین[۱]، پایین‌ترین[۲]، قیمت بسته شدن[۳] ،باز شدن[۴] و حجم معاملات[۵] هرروز بازار می‌باشد را به‌صورت یک فایل xls تولید می‌کند. با استفاده از این نرم‌افزار می‌توان نشانگرهای معروف را بر روی سهم موردنظر اعمال کرد و نتیجه را به‌صورت گرافیکی مشاهده کرد. منها به دلیل آنکه ما در این پژوهش نیاز به اجرای هم‌زمان و مقایسه تمام نشانگرها داریم، استفاده از این نرم‌افزار تنها جهت اخذ داده خام کاربرد دارد. و برای محاسبه نشانگرها ناچار به پیاده‌سازی تمام نشانگرها در محیط برنامه‌نویسی مانند متلب داریم.

نرم‌افزار متلب که اکثر پیاده‌سازی‌ها توسط آن انجام گرفت. اعم از پیاده‌سازی تمام نشانگرها، نرمالیزه کردن داده‌ها، کلاسه بندها به روش ماشین بردار پشتیبان و انتخاب ویژگی و درنهایت مقایسه نشانگرها.

[1] high
[2] low
[3] close
[4] open
[5] volume

مراجع

[1]. Pie-Chann Chang et al, "Trend Discovery in financial time series data using a case based fuzzy decision tree," *Expert Systems with Applications*, vol. 38.5, pp. 6070-6080, 2011.

[2]. Li-Ping Ni et al, "Stock trend prediction based on fractal feature selection and support vector machine," *Expert Systems with Applications*, vol. 38.5, pp. 5569-5576, 2011.

[3]. I. K. Fodor, "A survey of dimension reduction techniques," technical report, Lawrence Livemore National Laboratory, June 2002.

[4]. Yunyue Zhu, High Performance Data Mining in Time Series: Techniques and Case Studies, Ph.D. Dissertation, New York University, January 2004.

[5]. Lindsay I Smith, A tutorial on Principal Components Analysis, 2002.

[6]. M. Dash, H. Liu, Feature Selection for Classification. Intelligent Data Analysis 1:131-156, 1997.

[7]. Schlimmer, J.C., Efficiently inducing determinations: A complete and systematic search algorithm that uses optimal pruning. In: Proceedings of Tenth International Conference on Machine Learning, 284–290, (1993).

[8]. Kira, K. and Rendell, L.A., The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of Ninth National Conference on Artificial Intelligence, 129–134, 1992.

[9]. R. H. & Sap, M. N. M Khokhar, "Fuzzy decision tree for data mining of time series stock market database," *The fifth International conference for the Critical Assessment of Microarray Data Analysis*, 2004.

[10]. Jose Luis Aznarte et al, "Financial time series forcasting with a bio-inspired fuzzy model," *Expert Systems with Applications* , vol. 39.16, pp. 12302-12309, 2012.

[11]. Xu, L., Yan, P. and Chang, T., Best first strategy for feature selection. In: Proceedings of Ninth International Conference on Pattern Recognition, 706–708, 1988.

[12]. Bobrowski, L., Feature selection based on some homogeneity coefficient. In: Proceedings of Ninth International Conference on Pattern Recognition, 544–546, 1988.

[13]. Shaoning Pang et al, "Correlation-aided support vector regression for forex time series prediction," *Neural Computing and Applications*, vol. 20.8, pp. 1193-1203, 2011.

[14]. Ling-Jing Kao et al, "Integration of nonlinear independent component analysis and support vector regression for stock price forecasting," *Neurocomputing*, vol. 99, pp. 534-542, 2013.

[15]. Quinlan, J.R., C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, California, 1993.

[16]. Sheinvald, J., Dom, B. and Niblack, W., A modelling approach to feature selection. In: Proceedings of Tenth International Conference on Pattern Recognition, 1:535–539, June 1990.

[17]. Book.Portfolio Management Using Artificial Trading & Systems Based on Technical Analysis

[18]. Mucciardi, A.N. and Gose, E.E., A comparison of seven techniques for choosing subsets of pattern recognition. IEEE Transactions on Computers, C-20:1023–1031, September 1971.

[19]. Modrzejewski, M., Feature selection using rough sets theory. In: Proceedings of the European Conference on Machine Learning (P. B. Brazdil, ed.), 213–226, 1993.

[20]. Oliveira, A.L. and Vincentelli, A.S., Constructive induction using a non-greedy strategy for feature selection. In: Proceedings of Ninth International Conference on Machine Learning, 355–360, Morgan Kaufmann, Aberdeen, Scotland, 1992.

[21]. Liu, H. and Setiono, R., A probabilistic approach to feature selection—a filter solution. In: Proceedings of International Conference on Machine Learning, 319–327, 1996.

[22]. Lean Yu et al, "Evolving Least Squares Support Vector Machines for Stock Market Trend Mining," *Evolutionary Computation, IEEE Transactions*, vol. 13.1, pp. 87-102, 2009.

[23]. Devijver, P.A. and Kittler, J., Pattern Recognition: A Statistical Approach. Prentice Hall, 1982.

[24]. Caruana, R. and Freitag, D., Greedy attribute selection. In: Proceedings of Eleventh International Conference on Machine Learning, Morgan Kaufmann, New Brunswick, New Jersey, 28–36, 1994.

[25]. Doak, J., An evaluation of feature selection methods and their application to computer security. Technical report, Davis, CA: University of California, Department of

Computer Science, 1992.

[26]. Moore, A.W. and Lee, M.S., Efficient algorithms for minimizing cross validation error. In: Proceedings of Eleventh International Conference on Machine Learning, Morgan Kaufmann, New Brunswick, New Jersey, 190–198, 1994.

[27]. Domingos, P., Context-sensitive feature selection for lazy learners. Artificial Intelligence Review, 1996.

[28]. Queiros, C.E. and Gelsema, E.S., On feature selection. In: Proceedings of Seventh International Conference on Pattern Recognition, 1:128–130, July-Aug 1984.

[29]. Ichino, M. and Sklansky, J., Feature selection for linear classifier. In: Proceedings of the Seventh International Conference on Pattern Recognition, volume 1, 124–127, July–Aug 1984.

[30]. Ichino, M. and Sklansky, J., Optimum feature selection by zero-one programming. IEEE Trans. on Systems, Man and Cybernetics, SMC-14(5):737–746, September/October 1984.

[31]. Geoffrion, A.M., Integer programming by implicit enumeration and balas, method. SIAM Review, 9:178–190, 1967.

[32]. Foroutan, I. and Sklansky, J., Feature selection for automatic classification of non-gaussian data. IEEE Transactions on Systems, Man, and Cybernatics, SMC-17(2):187–198, 1987.

[33]. Liu, H. and Setiono, R., Feature selection and classification—a probabilistic wrapper approach. In: Proceedings of Ninth International Conference on Industrial and Engineering Applications of AI and ES, 284–292, 1996.

[34]. Book A to Z

[35]. George S. Atsalakis and Kimon P. Valavanis, "Surveying stock market forecasting techniques–Part II: Soft computing methods," *Expert Systems with Applications*, vol. 36.3, pp. 5932-5941, 2009.

[36]. BE Boser et al, "A training algorithm for optimal margin classifiers," *Proceedings of the fifth annual workshop on Computational learning theory. ACM*, pp. 144-152, 1992.

[37]. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002 (ISBN 981-238-151-1)

واژه‌نامه

نشانگر، شاخص فنی، اندیکاتور	indicator
آزمون، تست	Test
آموزش، ترین	Train
کلاس‌بندی، طبقه‌بندی	Classification
سهام	Stock
ماشین بردار پشتیبان	SVM, support vector machine
صحت	Accuracy

Abstract:

Analysis of financial market always attract a lot of attention from investors and researchers to have. The price of the stock market is complex and affected by many factors. So find the major factors affecting the stock market is very important and very difficult to predict the other hand, more recent articles researchers trying to predict the market with a series of features, usually with the aim of predicting the market price is, the results of it is unusable in the real market, so we decided to look from a different angle and applied research to the market to help traders do stock market, in which we seek to predict the future rather than the predicted price market that is divided into two categories descending and ascending, in this study we aim to test different parameters and find the influential factors that can predict future market uptrend or downtrend help us, things that we in this thesis we examined markers (indicators), which traders exploit them when predicting market trends. In order to predict the clustering support vector machine was used to identify factors affecting the rapper backward feature selection method was used to remove the characteristics of ineffective health and the properties remain effective. By implementing these algorithms on Shahrood cement stocks and Pars Oil could be the indicators most effective with respect to any stocks acquired time series.

Keywords: Support vector machine, predict stock trends, feature selection, indicator

University of shahrood

Faculty of E-Center

Computer Engineering – Artificial Intelligence

Developing a feature selection method for Iranian stock prediction

Javad khoddami

Supervisor:
Dr. Ali soleymani ivari

Advisor:
Eng. FarzanehAkbarzadeh

september 2015