فصلنامه علمی پژوهشی دانش سرمایه‌گذاری
دوره ۱۲ / شماره ٤ (پیاپی ٤٨) / زمستان ١٤٠٢
صفحه ۵۱۱ تا ۵۳۸

انتخاب بهینه سبدسهام با استفاده از الگوریتم های یادگیری ماشین

محمدباقر یزدانی خداشهری
دانشجوی دکتری، گروه حسابداری، واحد قائمشهر، دانشگاه آزاد اسلامی، قائمشهر، ایران

سیدحسین نسل موسوی
گروه حسابداری، واحد قائمشهر، دانشگاه آزاد اسلامی، قائمشهر، ایران، نویسنده مسئول مکاتبات

میرسعید حسینی شیروانی
گروه کامپیوتر، واحد ساری، دانشگاه آزاد اسلامی، ساری، ایران

تاریخ دریافت: ۹۹/۰۴/۲۹ تاریخ پذیرش: ۹۹/۰۵/۱۵

چکیده

انتخاب سبد سهام مناسب همواره از اساسی ترین مسائل سرمایه گذاران است. اساساً پیش بینی روند قیمت با استفاده از آنالیز فنی یا آنالیز اساسی انجام می شود. آنالیز فنی بر عملکرد بازار تمرکز دارد، در حالیکه تمرکز آنالیز اساسی مبتنی بر مکانیزم عرضه و تقاضا است و این سبب تغییر قیمت ها می شود. وجود راهکاری که بتواند رشد یا کاهش سهام را با استفاده از آن پیش بینی نماید، بعنوان یک نیاز اساسی در این تحقیق به آن پرداخته شده است. در پژوهش حاضر، به کمک دیتاستِ نظارت شده از راهکاری مبتنی بر الگوریتم های مجموعه راف و تحلیل سلسله مراتبی برای کاهش ویژگی و از الگوریتم های درخت تصمیم، ماشین بردار پشتیبان و شبکه بیزین برای پیش بینی استفاده شده است. این راهکار پیشنهادی با استفاده از زبان سی شارپ پیاده سازی و با راهکارهای مختلفی مقایسه شده و نتایج تحقیق نشان داده است که روش پیشنهادی با ۸۰ درصد دقت صحت پیش بینی و ۲۰ اشتباه در پیش بینی دارای بیشترین دقت و کمترین میزان اشتباه در میان روش های مورد مقایسه را دارد.

واژه‌های کلیدی: ماشین بردار پشتیبان، شبکه بیزین، درخت تصمیم بهبود یافته، مجموعه راف، انتخاب سهام.

۱- مقدمه

یکی از اهداف اصلی گردانندگان بازارهای پولی و مالی این است که هر کسی با هر سلیقه و هر مقدار و با انتخاب هر نوع دارایی بتواند وارد این بازارها شده و فرصت های مناسب سرمایه گذاری را تشخیص دهد و سود مناسبی کسب نماید. در همین راستا، از مهم ترین رسالت مهندسی مالی، طراحی ابزارهای متنوع مالی به منظور تنوع بخشیدن به بازارهای پولی و مالی می باشد[۲۲] [۳] [۱].

تحقیقات اخیر نشان می دهد که طراحی و ارائه مدل های قاعده محور برای زمان خرید و فروش امکانپذیر بوده و می توان با تولید قواعد معاملاتی، سیستم های توانمندی را برای پشتیبانی از تصمیم گیری های سرمایه گذاران، توسعه داد. انتخاب سبد سرمایه با تخصیص سرمایه محدود به تعدادی از دارایی های بالقوه سرمایه گذاری به منظور دستیابی به استراتژی سرمایه گذاری سودآور، در ارتباط است[۹]. سبد سرمایه، ترکیبی مناسب از سهام یا سایر دارایی ها است که یک سرمایه گذار آنها را خریداری می کند. هدف از تشکیل سبد سرمایه، توزیع ریسک سرمایه گذاری بین چند سهم است، بگونه ای که سود یک سهم بتواند ضرر سهام دیگر را جبران کند[۵] [۱]. در مساله بهینه سازی سبد سرمایه، هدف کمینه سازی ریسک و بیشینه سازی سود است. بنابراین ایجاد توازن بین ریسک و بازده، امری ضروری است. ایجاد این توازن به نظر ساده می باشد، اما در عمل روش های مختلفی برای تشکیل سبد سرمایه گذاری، استفاده می شود[۳] [۸].

دستیابی به رشد اقتصادی و ایجاد انگیزه جهت سرمایه گذاری، زمانی در یک کشور تسریع می گردد که آن کشور دارای بازارهای سرمایه فعال و قابل اعتماد باشد. وجود بازارهای بورس فعال همواره سرمایه گذاران متعددی را به تکاپو داشته و حرکت جریان سرمایه و منابع مالی را به بخش های مولد تسریع می نماید. یکی از کلیدهای موفقیت در بازار سرمایه برای سرمایه گذاران، اتخاذ رویکردی مناسب برای مدیریت سهام ساخته شده توسط آنها می باشد[۱][۲۲].

انتخاب سبد سرمایه با تخصیص سرمایه محدود به تعدادی از دارایی های بالقوه سرمایه گذاری به منظور دستیابی به استراتژی سرمایه گذاری سودآور، در ارتباط است. سبد سرمایه، ترکیبی مناسب از سهام یا سایر دارایی ها است که یک سرمایه گذار آنها را خریداری می کند. هدف از تشکیل سبد سرمایه، توزیع ریسک سرمایه گذاری بین چند سهم است به گونه ای که سود یک سهم بتواند ضرر سهام دیگر را جبران کند. در مساله بهینه سازی سبد سرمایه، هدف کمینه سازی ریسک و بیشینه سازی سود است. بنابراین ایجاد توازن بین ریسک و بازده امری ضروری است[۹][۵][۳].

ایجاد این توازن به نظر ساده می باشد، اما در عمل روش های مختلفی برای تشکیل سبد سرمایه گذاری، استفاده می شود. هدف از حل مدل های انتخاب سبد سرمایه گذاری، ارائه مجموعه ای جواب ها به تصمیم گیرندگان جهت انتخاب سبد سرمایه مورد نظر است.

اولین مطالعه در رابطه با مساله انتخاب پورتفولیو، بکارگیری مفهوم مجموعه کارا است که توسط مارکویتز(۱۹۵۲) ارائه شد[۵]. مطالعه او که مبنای تئوری پورتفولیوی مدرن به شمار می رود، انتخاب پورتفولیو را در قالب یک مساله بهینه سازی میانگین- واریانس با دو معیار اساسی: ۱- حداکثر سازی سود (میانگین بازده مورد

انتظار) و ۲- حداقل سازی ریسک (واریانس بازده مورد انتظار) مورد توجه قرار داده است. بدیهی است که یک پورتفوی مطلوب با توجه به بده بستان ریسک و بازده مورد انتظار تعیین می شود[۱۲][۱].

با وجود اینکه مدل میانگین- واریانس مارکویتز، مبنای تئوری پورتفولیوی مدرن به شمار می رود، به دلیل اینکه مدلی ساده سازی شده با مفروضات غیر واقع گرایانه است، استفاده عملی چندانی ندارد. از نقطه نظر عملی، سرمایه گذاران معمولا با محدودیت های الزام آوری روبه رو هستند. هر چقدر محدودیت های عملی به منظور توسعه مدل در مدلسازی افزایش یابد، حل مدل مشکل تر می شود. بسیاری از محققان از انواع تکنیک ها برای حل مساله انتخاب پورتفولیوی مقید استفاده کرده اند[۱۸]. هدف از حل مدل های انتخاب سبد سرمایه گذاری، ارائه مجموعه ای جواب ها به تصمیم گیرندگان جهت انتخاب سبد سرمایه مورد نظر است[۶].

سوال اساسی در این پژوهش این است که چگونه می توان با استفاده از از الگوریتم های ماشین، تغییرات قیمت سهام را پیش بینی و با استفاده از آن سبد بهینه سهام را شناسایی کرد؟

در این تحقیق با استفاده از راهکاری مبتنی بر مجموعه راف و تحلیل سلسله مراتبی برای تعیین ویژگی های موثر و بکارگیری الگوریتم های درخت تصمیم بهینه شده، ماشین بردار پشتیبان و شبکه بیزین بصورت ترکیبی برای پیش بینی انتخاب و یا عدم انتخاب سهام استفاده شده تا بتوان بوسیله آنها میزان تغییرات قیمت هر یک از سهام را پیش بینی و سهامی که دارای بیشترین تغییرات از لحاظ افزایش قیمت هستند را شناسایی و در قالب سبد سهام ارائه و بدین شکل بهترین سبد سهام انتخاب و به تبع آن بیشترین سود نیز کسب گردد.

مبانی نظری و ادبیات پژوهش

الگوریتم های یادگیری ماشین که در این تحقیق از آنها استفاده شده شامل، ماشین بردار پشتیبان، شبکه بیزین، درخت تصمیم و مجموعه راف و تحلیل سلسله مراتبی است؛ که ذیلاً تشریح شده اند:

ماشین بردار پشتیبان

ماشین بردار پشتیبان یکی از روش‌های یادگیری با نظارت است که از آن برای طبقه‌بندی و رگرسیون استفاده می‌کنند[۲۳]. این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای طبقه‌بندی نشان داده‌است. 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 به وسیله روش‌های QP که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد. قبل از تقسیمِ خطی برای اینکه ماشین بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله 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 مورد نظر به فرم دوگانی آن که در آن به جای 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 از جمله هسته‌های نمایی، چندجمله‌ای و سیگموید می‌توان استفاده نمود[۲۵].

شبکه بیزین

شبکه های بیزین که با نام شبکه های اعتقاد (باور) هم شناخته می شوند، متعلق به خانواده مدل های گرافیکی احتمالاتی هستند[۲۶]. این ساختارهای گرافیکی برای نشان دادن اطلاعات در یک حوزه دارای عدم قطعیت به کار می روند[۲۷]. به طور خاص هر گره در گراف نشان دهنده یک متغیر تصادفی است و شاخه ها (کمان) وابستگی های احتمالاتی بین متغیرها را نشان می دهند. این وابستگی های شرطی غالبا به وسیله روش های آماری و احتمالاتی مشخص ارزیابی می شوند. شبکه های بیزین اصولی از نظریه گراف، نظریه احتمالات، علوم کامپیوتر و آمار را با هم ترکیب می کنند[۲۶].

شبکه بیزی یک گراف جهت‌دار است که رئوس آن شامل اطلاعات مقادیر احتمالات شرطی هستند. بطور دقیق‌تر این شبکه شامل اجزا و خصوصیات زیر است[۲۷]:

- یک مجموعه از متغیرهای تصادفی، مجموعه رئوس گراف را تشکیل می‌دهند که این متغیرها می‌توانند گسسته یا پیوسته باشند.
- یک مجموعه از یال های جهت‌دار که اگر یک یال از راس به راس باشد، را والد می‌نامیم.
- هر گره، یک توزیع احتمال شرطی دارد که تاثیر گره‌های والد بر روی این گره را بصورت عددی نشان می‌دهند.
- گراف هیچ دور جهت‌داری ندارد و در واقع، یک گراف بدون دور جهت‌دار است.

ساختار شبکه نشان دهنده وابستگی‌های شرطی در قلمرو است. بصورت شهودی، معنی یک یال از به وجود تاثیر مستقیم بر و یا وابستگی مستقیم به است. باید توجه داشت که تعیین این وابستگی‌های مستقیم برای یک فرد خبره قلمرو کار مشکلی نمی‌باشد و به همین دلیل معمولاً در صورت وجود فرد خبره تعیین ساختار شبکه آن چنان سخت نمی‌باشد. پس از تعیین ساختار، تعیین توزیع شرطی مربوط به گره‌ها، ساختمان داده شبکه بیزی را کامل می‌کند و با استفاده از آن می‌توان توزیع توام کامل را بدست آورد[۲۸].

درخت تصمیم

ساختار درخت تصمیم در یادگیری ماشین، یک مدل پیش بینی کننده می باشد که حقایق مشاهده شده در مورد یک پدیده را به استنتاج هایی در مورد مقدار هدف آن پدیده نقش می کند. تکنیک یادگیری ماشین برای استنتاج یک درخت تصمیم از داده ها، یادگیری درخت تصمیم نامیده می شود که یکی از رایج ترین روش های داده کاوی است[۱۲].

هر گره داخلی، متناظر یک متغیر و هر کمان به یک فرزند، نمایانگر یک مقدار ممکن برای آن متغیر است. یک گره برگ، با داشتن مقادیر متغیرها که با مسیری از ریشه درخت تا آن گره برگ بازنمایی می شود، مقدار پیش بینی شده متغیر هدف را نشان می دهد. یک درخت تصمیم ساختاری را نشان می دهد که برگ ها نشان دهنده دسته بندی و شاخه ها ترکیبات فصلی صفاتی که منتج به این دسته بندی ها را بازنمایی می کنند[۱۳]. یادگیری یک درخت می تواند با تفکیک کردن یک مجموعه منبع به زیرمجموعه هایی براساس یک تست مقدار صفت انجام شود. این فرآیند به شکل بازگشتی در هر زیرمجموعه حاصل از تفکیک تکرار می شود. عمل بازگشت

زمانی کامل می شود که تفکیک بیشتر سودمند نباشد یا بتوان یک دسته بندی را به همه نمونه های موجود در زیرمجموعه بدست آمده اعمال کرد[۱۴].

درختان تصمیم قادر به تولید توصیفات قابل درک برای انسان، از روابط موجود در یک مجموعه داده ای هستند و می توانند برای وظایف دسته بندی و پیش بینی بکار روند. این تکنیک به شکل گسترده ای در زمینه های مختلف همچون تشخیص بیماری دسته بندی گیاهان و استراتژی های بازاریابی مشتری بکار رفته است.

این ساختار تصمیم گیری می تواند به شکل تکنیک های ریاضی و محاسباتی که به توصیف، دسته بندی و عام سازی یک مجموعه از داده ها کمک می کنند نیز معرفی شوند[۱۷].

انواع صفات در درخت تصمیم به دو نوع صفات دسته ای و صفات حقیقی بوده که صفات دسته ای، صفاتی هستند که دو یا چند مقدار گسسته می پذیرند (یا صفات سمبلیک) درحالی که صفات حقیقی مقادیر خود را از مجموعه اعداد حقیقی می گیرند[۱۸].

مجموعه راف

بسیاری از مفاهیم و تئوری های عدم قطعیت نظیر مجموعه های فازی ، سیستم های خاکستری و مجموعه های راف، در گذشته معرفی شده و در سال های اخیر ابزارهای ریاضی مبتنی بر آن ها با سرعت بالایی توسعه یافته اند. هریک از این رویکرد ها ، مفاهیم خاص خود را داشته و دارای ویژگی های منحصر به خود می باشد. به عنوان مثال ، تئوری کلاسیک به دنبال تحلیل داده های احتمالی یا قطعی بوده و تئوری فازی ، محاسبات نرم را اساس کار خود قرار داده است[۱۹]. تئوری خاکستری به کنترل سیستم ها در شرایط کمبود داده ها و اطلاعات ناکامل پرداخته و تئوری راف ، تقریب و استدلال درباره داده ها را بدنبال دارد . داده هایی که از دنیای واقعی اخذ می گردند معمولا شامل تمامی انواع نویزها بوده و عدم قطعیت بسیار و اطلاعات غیر کامل فراوانی به همراه دارند. روش های سنتی برخورد با این عدم قطعیت نظیر تئوری فازی ، تئوری گواه، تئوری احتمالات و نظایر آن، به اطلاعات اضافی مانند توزیع احتمال و تابع عضویت نیازمند هستند . به بیان دیگر ، کار با این سیستم ها به دلیل حجم بالایی از داده ها مشکل است ، از این رو به کارگیری سایر تئوری ها نظیر تئوری مجموعه های راف می تواند در این راه کمک کننده باشد[۲۰]. مجموعه راف ابزاری قابل استفاده از شرایط ابهام و عدم قطعیت است که اولین بار توسط پاولاک (۱۹۸۲) ارائه شد. این تئوری راف در زمینه -های مختلفی مورد استفاده قرار می گیرد از جمله تجزیه و تحلیل تصمیم گیری، سیستم های پشتیبان تصمیم. بعد از آقای پاولاک، سه محقق دیگر بنام-های ژای ، خو و ژانگ در سال ۲۰۰۸ اعداد راف را ارائه کردند. یک عدد راف دارای حد پایین (L)، حد بالا (U) و حد میانی که به فاصله مرزی راف مشهور است تشکیل شده است. اعداد راف در مسائلی استفاده می- شود که نظرات خبرگان در آن دخیل هستند و به نوعی باعث ایجاد عدم قطعیت و ابهام بشود [۱۱].

تحلیل سلسله مراتبی

فرایند واکاوی سلسله مراتبی یکی از روش‌های تصمیم‌گیری است. واژه AHP مخفف عبارت Analytical Hierarchy process به معنی فرایند تحلیل سلسله مراتبی است.انتخاب سنجه‌ها یا criterion بخش اول واکاوی AHP است. سپس براساس سنجه‌های شناسایی شده نامزدها ارزیابی می‌شوند. واژه گزینه‌ها یا نامزدها هم معنای

واژه alternative یا candidates بوده و به جای هم بکار روند. علت سلسله مراتبی خواندن این روش آن است که ابتدا باید از اهداف و راهبردهای سازمان در راس هرم آغاز کرد و با گسترش آن‌ها سنجه‌ها را شناسایی کرد تا به پایین هرم برسیم. این روش یکی از روش‌های پرکاربرد برای رتبه‌بندی و تعیین اهمیت عوامل است که با استفاده از مقایسات زوجی گزینه‌ها به اولویت بندی هر یک از معیارها پرداخته می‌شود. چنانچه گزینه‌ها زیاد باشد تشکیل ماتریس مقایسات زوجی کار دشواری است [۱۰].

مفهوم سرمایه گذاری و بورس

بطورکلی سرمایه گذاری عبارت است از تبدیل وجوه مالی به یک یا چند نوع دارائی دیگر و نگهداری آن برای مدتی در زمان آینده، به همراه پذیرفتن ریسک (خطر) مشخص یا نامشخص، برای کسب سود در آینده می باشد. به مجموعه هایی از رفتار ها، فرآیندها و قوانین که سرمایه گذار را به سمت بهترین سرمایه گذاری هدایت می کند، استراتژی های سرمایه گذاری می گویند[۳].

عوامل مختلفی مانند اهداف بلند مدت و کوتاه مدت، آستانه ریسک پذیری و اهداف شخصی سرمایه گذار در بکارگیری استراتژی های سرمایه گذاری نقش دارند[۵].

مروری بر پیشینه پژوهش

میرعلوی و پورزمانی(۱۳۹۸)، در مطالعه ای با ارائه مدلی جهت پیش بینی قیمت سهام با استفاده از روش های فرا ابتکاری و شبکه های عصبی پرداخته اند. مدل پیشنهادی در این پژوهش یک سیستم دو سطحی از شبکه های عصبی پرسپترون چندلایه و چندین شاخص برای پیش بینی استفاده شده است. از الگوریتم بهینه سازی ملخ برای انتخاب بهترین نمونه استفاده شده است. نتایج حاصل از پژوهش نشان داده که مدل پیشنهادی توانسته با خطای پیش بینی پایین تری نسبت به دیگر مدل ها عمل کند[۷].

خنده خوش(۱۳۹۵) با در نظر گرفتن عوامل مؤثر در پیش‌بینی شاخص قیمت بورس تهران با بهبود الگوریتم بهینه‌سازی ملخ در انتخاب بهترین نمونه‌ها در مدل آموزش چندتایی شبکه عصبی پرداخته است. به دلیل پیچیدگی بازار بورس و حجم بالای اطلاعات مورد پردازش، اغلب استفاده از یک سیستم ساده برای پیش‌بینی نتایج خوبی به همراه ندارد. به همین دلیل محققان با ارائه مدل‌های ترکیبی سعی در ارائه سیستمی با پیچیدگی کمتر و کارآیی و دقت بیشتر کرده‌اند. در اکثر مدل‌های پیش‌بینی کننده، سیستم فقط با استفاده از اطلاعات یک شاخص به پیش‌بینی پرداخته، ولی در مدل پیشنهادی یک سیستم دوسطحی از شبکه‌های عصبی پرسپترون چندلایه پیشنهاد می‌شود و از چندین شاخص برای پیش‌بینی استفاده می‌شود و همچنین برای آموزش بهتر شبکه عصبی و در نتیجه بهبود نتایج به‌دست‌آمده، از الگوریتم بهینه‌سازی ملخ برای انتخاب بهترین نمونه‌ها برای آموزش شبکه عصبی استفاده‌شده است و نتایج به‌دست‌آمده نشان می‌دهد که مدل پیشنهادی توانسته با خطای پیش‌بینی پایین‌تری نسبت به دیگر مدل‌ها عمل کند [۸].

یانگ سین و همکاران (۲۰۱۳) در مطالعه ای به بررسی الگوریتم pso در برش سهام نامنظم پرداختند. مشکلات برش سهام (CSP) در بسیاری از صنایع تولیدی بوجود می آیند که در آن ورق های سهام بزرگ باید به

قطعات کوچکتر برسد. در این تحقیق یک رویکرد تودرتو نامنظم برای مسئله برش دو بعدی طراحی شده است. یک روش اکتشافی مبتنی بر الگوریتم بهینه سازی (PSO) برای مسئله برش دو بعدی برش شکل، که در آن PSO برای یافتن راه حل بهینه استفاده می شود، ارائه می دهد. علاوه بر این، رویکرد پیشنهادی یک روش تقریبی شبکه ای با استراتژی قرار دادن اکتشافی پایین سمت چپ را برای تخصیص موارد نامنظم ترکیب می کند. در این پژوهش، رویکرد پیشنهادی با استفاده از ۱۵ معیار بازنگری و ارزیابی می شود. عملکرد نشان دهنده اثربخشی و کارایی این رویکرد در حل مشکلات برش سهام نامنظم است [۲۴].

چن و همکاران(۲۰۱۵) در مقاله ای با عنوان الگوریتم ژنتیک رابطه ای به همراه جهش هدایت شده برای بهینه سازی پرتفوی مقیاس بزرگ که از مهمترین تحقیقات اخیر در زمینه موضوع تحقیق می باشد، از الگوریتم ژنتیک رابطه ای به همراه یک عملکر جدید ارائه می نماید که جهش خوانده می شود. برای انتخاب موثرترین پرتفوی الگوریتم ژنتیک رابطه ای، ضرایب همبستگی بین برندهای سهام را به عنوان قدرت در نظر می گیرد که نشان دهنده رابطه بین گره ها در هر یک از الگوریتم ژنتیک رابطه ای می باشد. جهش هدایت شده پرتفوی جدید را مطابق با مقدار میانگین ضرایب همبستگی بین سهام تولید می کند که به معنی قابلیت بهره برداری از تکامل الگوریتم ژنتیک رابطه ای است. نتایج این تحقیق حاکی از آن است که راهکار الگوریتم ژنتیک رابطه ای با جهش هدایت شده، موفق بوده است و پرتفوی بدست آمده در محدوده یا نزدیک به محدوده قابل قبول می گنجد [۱۶].

ایستون و پیتر(۲۰۱۶)، در مطالعه ای دریافتند که در برش های مقطعی از سهام در بورس نیویورک که از سال ۱۹۲۶ به طور مداوم معامله شده اند، نسبت قیمت به سود تقسیمی (P/D) یکی از قویترین پیش بینی کننده های ارزش فعلی تغییرات سودهای آتی است. بنابراین، نسبت (P/D) پیش بینی هایی از تغییرات بلند مدت در سود های آتی ارائه می کند. این بدین معنی نیست که حباب های با اهمیتی در قیمت تک تک سهام وجود ندارد. به همان اندازه که حباب از دریا بیرون می آید، تغییر قابل پیش بینی در سود سهام شرکت ها اتفاق می افتد. تعداد زیادی از این تغییرات قابل پیش بینی در شرکت هایی است که سال های متمادی هیچ پرداخت سودی نداشتند، ولی سرمایه گذار به درستی در می یابد که سود پرداخت خواهد شد [۲۱].

روش پیشنهادی

روشی که در این تحقیق استفاده شده است یک روش ترکیبی می باشد که بتوان به این شکل دقت روش را بسیار افزایش داد. در اینجا از الگوریتم بیزین، SVM و درخت تصمیم ID3 استفاده شده است. با استفاده از این راهکار پیشنهادی برای هر یک از این الگوریتم ها وزنی در نظر گرفته می شود و برای محاسبه و پیش بینی مقدار محاسبه شده توسط هر یک از این الگوریتم ها در وزن آن الگوریتم ضرب می شود و در انتها نتیجه واقعی بدست می آید که دارای دقت بالاتری خواهد بود زیرا از فواید هر دو این الگوریتم ها استفاده نموده است.

روش پیشنهادی دارای مراحل مختلفی می باشد که به ترتیب عبارتند از:

- پیش پردازش داده ها و انتقال داده ها
- انتخاب ویژگی های موثر با استفاده از مجموعه راف و AHP

- ساخت درخت تصمیم گیری
- آموزش و محاسبه اوزان الگوریتم های درخت تصمیم، ماشین بردار پشتیبان و الگوریتم بیزین

فلوچارت کلی روش پیشنهادی را می توان در شکل (۲) مشاهده نمود.

شکل (۲): مدل پیشنهادی تحقیق

پیش پردازش داده ها

در ابتدا مجموعه داده جمع آوری شده و به آماده سازی و پیش پردازش داده ها پرداخته می شود. در آماده سازی و پیش پردازش داده ها از روش های مختلفی استفاده می شود. اول این که برخی ویژگی ها دارای مقادیر منحصر به فرد هستند. این ویژگی ها نمی توانند دانش مفیدی را در مجموعه داده ایجاد کنند. لذا این مجموعه ویژگی ها باید از داده ها حذف شوند. به طور نمونه می توان به ویژگی نام و نام خانوادگی اشاره نمود. همچنین ممکن است برخی تراکنش ها دارای مقادیر مفقود فراوان باشند. لذا این تراکنش ها نیز باید از مجموعه داده ها حذف شوند. از طرفی ممکن است، مقادیر برخی ویژگی ها دارای مقادیر نویز و مفقود باشند لذا این مقادیر نیز باید در مجموعه داده اصلاح شوند. مرحله بعدی نوبت به استفاده از ابزار کشف آنومالی پرداخته می شود. داده هایی که در نقاط خارج از قانون مجموعه داده قرار دارند شناسایی شده و حذف می شوند. برای اینکه بتوان روی داده ها به عنوان ورودی کار کرد می بایست ویژگی هایی را از آن ها استخراج نمود. به طور معمول پیش از انتخاب و استخراج ویژگی ها، برخی عملیات پیش پردازش بر روی داده ها انجام می شود.

انتقال داده ها

در این قسمت داده ها در دامنه های درست قرار می گیرند. بدین معنا که داده ها باید به رنج هایی که در سیستم مشخص شده است منتقل شوند و داده های خارج از رنج، داده های مشکل دار بوده و می بایست حذف شوند. داده های می بایست در محدوده درست قرار بگیرند بدین معنا که برای مثال اگر فیلد سن وجود داشته باشد فردی که محدوده سنی بین ۵۵ تا ۷۰ دارد می بایست در سیستم بصورت خیلی پیر شود که این قسمت بصورت اتوماتیک از روی مجموعه داده ها تکمیل می شود.

انتخاب ویژگی های موثر با استفاده از مجموعه راف و تحلیل سلسله مراتبی

مجموعه راف ابزاری قابل استفاده از شرایط ابهام و عدم قطعیت است که اولین بار توسط پاولاک۱ (۱۹۸۲) ارائه شد. این تئوری راف در زمینه -های مختلفی مورد استفاده قرار می گیرد از جمله تجزیه و تحلیل تصمیم گیری، سیستم های پشتیبان تصمیم. بعد از آقای پاولاک، سه محقق دیگر بنام های ژای۲، خو۳ و ژانگ۴ در سال ۲۰۰۸ اعداد راف را ارائه کردند. یک عدد راف دارای حد پایین (L)، حد بالا (U) و حد میانی که به فاصله مرزی راف۵ مشهور است تشکیل شده است. اعداد راف در مسائلی استفاده می شود که نظرات خبرگان در آن دخیل هستند و به نوعی باعث ایجاد عدم قطعیت و ابهام بشود.

فرض کنید که در یک مجموعه تصمیم-گیری مجموعه Uشامل تمام اعضای مجموع باشد. Y یک عضو دلخواه از مجموعه U و R یک مجموعه از t کلاس است که تمام اعضای U را پوشش می دهد. اگر این کلاس ها به صورت ترتیبی، همانند G1<G2<…<Gt باشند آنگاه حدهای پایین، بالا و ناحیه مرزی از کلاس G به صورت رابطه (۱) تعریف می شود.

فرمول ۱)

$$\underline{Apr}(G_q) = \bigcup\{Y \in U \mid R(Y) \leq G_q\}$$
$$\overline{Apr}(G_{q)=} \bigcup\{Y \in U \mid R(Y) \geq G_q\}$$
$$Bnd(G_q) = \bigcup\{Y \in U \mid R(Y) \neq G_q\} =$$
$$\{Y \in U \mid R(Y) \succ G_q\} \cup \{Y \in U \mid R(Y) \prec G_q\} \underline{Apr}(G_q) = \bigcup\{Y \in U \mid R(Y) \leq G_q\}$$

سپس این کلاس G می تواند به صورت یک عدد راف در حدهای پایین و بالا به صورت رابطه (۲) ارائه شود.

[1] Pawlak
[2] Zhai
[3] Khoo
[4] Zhong
[5] Rough boundary interval

فرمول ۲)

$$\underline{Lim}(G_q) = \frac{1}{M_L}\sum R(Y)|Y \in \underline{Apr}(G_q)$$
$$\overline{Lim}(G_q) = \frac{1}{M_L}\sum R(Y)|Y \in \overline{Apr}(G_q)$$
$$RN(G_q) = \left[\underline{Lim}(G_q), \overline{Lim}(G_q)\right]$$

همچنین فاصله مرزی راف بصورت شکل (۳) محاسبه می شود این فاصله مرزی ابهام را بیان می کند به طوریکه هر چقدر این عدد بزرگتر باشد نشان دهنده ابهام بیشتر است و اگر عدد کوچکتر باشد نشان دهنده دقت بیشتر است.

شکل (۳): محاسبه محدده در مجموعه راف

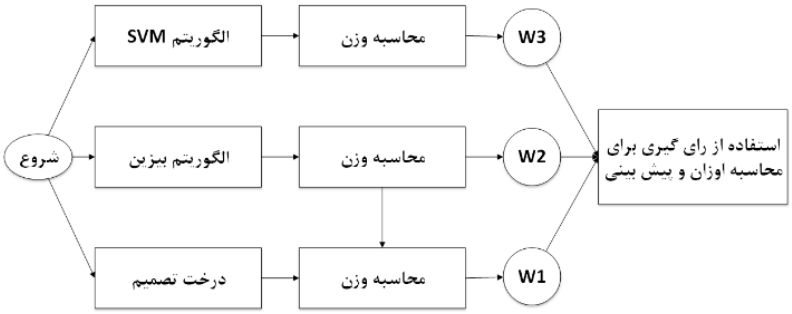
روش AHP (فرایند تحلیل سلسله مراتبی) از روش های پر کاربرد در تصمیم گیری چند معیاره است که هدف آن محاسبه وزن معیارها و گزینه- های پژوهش تحت یک مدل سلسله مراتبی است. در این مدل ابتدا مقایسات زوجی تشکیل شده و در اختیار خبرگان قرار داده می شود تا بر اساس طیف ۱ تا ۹ نظرات خود را نسبت به مقایسه دو به دوی معیارها بیان کنند. برای استفاده از اعداد راف در روش AHP (rough AHP) به طریق زیر عمل می کنیم.

۱) ابتدا مقایسات زوجی خبره ها را از نظر نرخ ناسازگاری بررسی کرده و چنانچه نرخ ناسازگاری کمتر از ۰.۱ باشد یعنی مقایسه زوجی سازگار است و در صورتیکه بزرگتر از ۰.۱ باشد باید اعداد مقایسه زوجی اصلاح شود.
۲) ایجاد اعداد راف از اعداد خبره ها با استفاده از روابطی که در تئوری گفته شد.
۳) محاسبه وزن فاصله ای معیارها با استفاده از روش میانگین هندسی

آموزش و محاسبه اوزان الگوریتم های درخت تصمیم، ماشین بردار پشتیبان و شبکه بیزین

در ابتدای کار در صدی از مجموعه داده های مورد استفاده برای محاسبه آموزش و محاسبه وزن مورد استفاده قرار می گیرد. در این قسمت قصد استفاده از یک الگوریتم ترکیبی می باشد که از دو الگوریتم درخت تصمیم و ماشین

بردار پشتیبان استفاده می کند. این دو الگوریتم هر کدام سهمی از جواب نهایی را خواهند داشت که بدین شکل دقت سیستم افزایش می یابد. می توان نمایی از این مرحله را در شکل (۴) مشاهده نمود.



شکل (۴): معماری استفاده شده در محاسبه وزن

همانطور که در شکل (۴) مشاهده می شود الگوریتم پیشنهادی در ابتدا با استفاده از یک مجموعه داده به محاسبه وزن می پردازد و این محاسبه بدین ترتیب می باشد که هر الگوریتم با استفاده از ۷۰ درصد مجموعه داده های موجود آموزش می بیند و با استفاده از ۳۰ درصد باقیمانده مورد آزمون قرار می گیرد در نهایت با توجه به تعداد جواب های صحیح امتیاز و یا وزنی به آن تعلق می گیرد تا اینکه با استفاده از وزن تعلق توان در مرحله بعدی وزنی را برای خروجی هر الگوریتم در نظر گرفت همانطور که در معماری نیز می توان مشاهده نمود بعد از محاسبه وزن که بصورت تقسیم تعداد جواب های درست به تعداد کل جواب های حدس زده شده می باشد، می توان میزان تاثیر گذاری هر کدام از این الگوریتم را در خروجی نهایی بهتر تشخیص داد. در این روش بعد از محاسبه وزن ها، به ازای هر رکورد ، پیش بینی ای توسط ماشین بردار پشتیبان، شبکه بیزین و توسط درخت تصمیم بهینه صورت می گیرد که مقدار پیش بینی شده می بایست در وزن آن الگوریتم ضرب شود و خروجی نهایی پیش بینی الگوریتم برابر است با جمع نتایج هر یک از الگوریتم ضرب در وزن آن الگوریتم که بدین صورت نتیجه نهایی بدست می آید و دسته بندی درست صورت می گیرد. در واقع در این قسمت از روش رای گیری (Voting) استفاده می شود.

ساخت درخت تصمیم گیری

در این قسمت از درخت تصمیم گیری استفاده می شود. درخت تصمیم گیری، درختی است که هر شاخه از آن به عنوان یک انتخاب می باشد. بدین معنی برای رفتن از گره ریشه به گره پایین تر می توان از شاخه هایی که به آن گره متصل هستند یکی انتخاب شود. در انتها هر یک از گره های انتهایی یا اصطلاحا گره برگ تصمیمی را بازگو می کند. هر کدام از شاخه ها تا رسیدن به برگ دارای سناریویی می باشد که موجب اتخاذ یک تصمیم می شود. در این پژوهش از مدل پیشنهادی مبتنی بر درخت تصمیم ID3 بهبود یافته استفاده شده است که این بهبود موجب سرعت عمل بالای آن شده است. درخت ID3 یک درخت تصمیم گیری می باشد که دارای یادگیری نیز

می باشد و اولین بار توسط راس کوینلن[1] مطرح شد. ایده الگوریتم ID3، ساخت درخت تصمیم گیری بالا به پایین می باشد که انتخاب گره در آن به وسیله جستجوی حریصانه از میان مجموعه ای از صفت ها می باشد. در اینجا ما برای اینکه قادر باشیم تا مفیدترین صفت را از میان صفات بیابیم که در کلاسه بندی مفید تر باشد از الگویی بخصوص استفاده نمودیم. برای اینکه بتوان کلاسه بندی مفیدی را برای مجموعه یادگیری انجام داد، می بایست تعداد سوالات را کاهش داد یا می توان گفت می بایست عمق درخت تصمیم گیری را کاهش داد. از اینرو در این قسمت نیاز به تابعی است که قادر باشد تا متعادل ترین تقسیم را انجام دهد که در این صورت عمق درخت بسیار کاهش می باید و گره ها بصورت متعادل در درخت تقسیم می شوند.

جدولی را در نظر بگیرید که دارای صفات و کلاسی از صفات نیز می باشد. در صورتی به این جدول همگن[2] گفته می شود که تنها شامل یک کلاس باشد. اگر یک جدول دارای چندین کلاس باشد، در این حالت به آن ناهمگن[3] گویند. توابع زیادی همچون آنتروپی، gini index و classification error برای سنجش میزان همگن پذیری وجود دارند. در این میان در اینجا از آنتروپی[4] استفاده شده است.

$$\text{Entropy} = \sum_{j} - p_j \log_2 p_j$$

فرمول ۳)

آنتروپی یک جدول صفر است زیرا احتمال آن مقداری برابر یک است(تنتها دارای یک کلاس باشد). آنتروپی زمانی به بیشترین مقدار خود می رسد که تمامی کلاس های موجود در جدول دارای احتمالی برابر باشند. آنتروپی را می توان به نوعی معیاری برای سنجش بی نظمی در نظر گرفت هر چه مجموعه منظم تر باشد و دارای گوناگونی کمتری باشد آنگاه آنتروپی آن کمتر است و به نوعی بی نظمی آن نیز کمتر است و برعکس. البته در اینجا چون ما در مرحله قبل یک کلاسه بندی ابتدایی را انجام دادیم تقریبا بی نظمی نیز پایین می باشد و این خود باعث سرعت عمل بالاتر روش پیشنهادی ما می شود زیرا این قضیه باعث می شود که عمق درخت تصمیم گیری کم شود و هر چه عمق این درخت کمتر باشد، سرعت تصمیم گیری نیز بیشتر می شود.

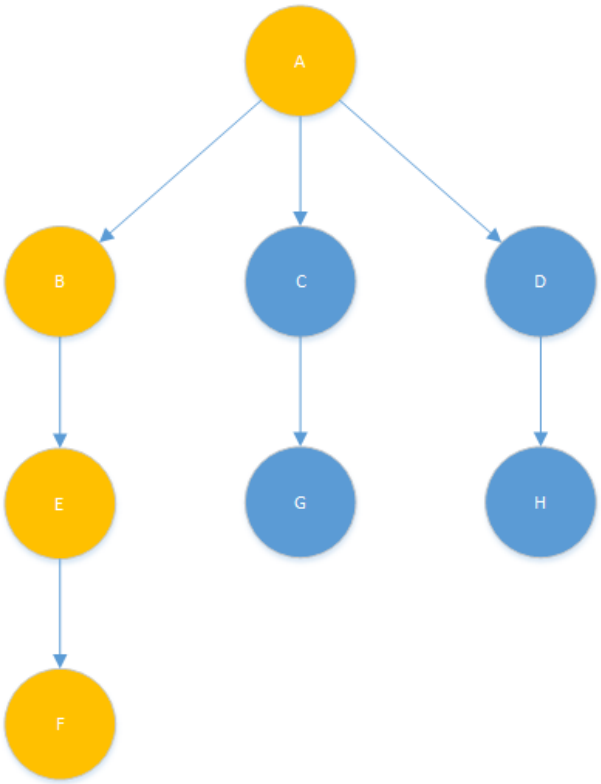
در این قسمت ما از آنتروپی استفاده کردیم تا مقدار بی نظمی را برای هر صفت از جدول (۱) بدست آوریم. برای اینکه بتوانیم صفتی را در درخت تصمیم گیری انتخاب کنیم که در رتبه بالاتری از بقیه صفت ها باشد به نوعی دارای اهمیت بالاتری از بقیه صفت ها باشد از فرمول (۳) استفاده کردیم. با توجه به این فرمول ما آنتروپی همه صفات را در مجموعه S محاسبه می کنیم و مقدار صفت مجموعه A را از آن می کاهیم. مجموعه A، مجموعه همه صفات اتخاب شده از پدر تا به اینجا در یک مسیر خاص می باشد.

$$G(S, A) = Entropy(S) - \sum_{v \in Values(A)} Entropy(v)$$

فرمول ۴)

[1] Ross Quinlan
[2] homogenous
[3] heterogeneous
[4] Entropy

برای درک بهتر فرمول (۴) می توان شکل (۵) را مشاهده نمود.



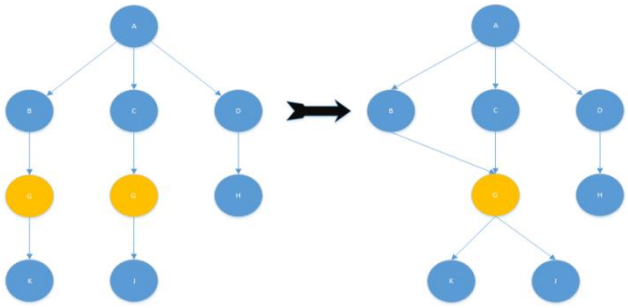
شکل (۵): مثالی برای انتخاب صفت ها

همانطور که در شکل (۵) قابل مشاهده است، ما می بایست آنتروپی تمامی صفات را از آنتروپی صفات انتخاب شده تا به اینجای مسیر بکاهیم(یعنی باید $G(S,F)=E(S)-(E(A)+E(B)+E(E)+E(F))$ را بدست آورد). البته باید توجه شود که ما می بایست در مجموعه A صفاتی که تا به اینجای کار استفاده شده است به علاوه صفتی که می خواهیم قرار دهیم را محاسبه نماییم. بعد از اینکار از بین این مجموعه صفات باقیمانده که برای هر کدام فرمول (۲) را محاسبه نمودیم، صفتی را که دارای G بیشتری است را انتخاب نماییم. در این حالت اگر دو صفت دارای G برابر بودند که احتمال این پیشامد نیز کم نیست، می بایست به گره دو یا هر تعداد صفت که دارای بیشترین مقدار G هستند و باهم برابر نیز می باشند به گره مربوطه بیفزاییم یعنی اگر برای مثال در گره ای دو صفت دارای G برابر بودند، آنگاه این دو به گره مربوطه دو فرزند می افزاییم و هر کدام از این صفت ها به عنوان یک فرزند این گره در نظر گرفته می شوند و بعد از این کار روند الگوریتم را برای هر یک از این گره ها ادامه می دهیم. برای مثال در

شکل (۵) می توان مشاهده نمود که مقدار G برای صفت های B، C و D برابر است از این رو همه این صفت ها در یک سطح قرار گرفته اند.
این کار باعث می شود تا صفت هایی که دارای آنتروپی بیشتری هستند را بیابیم چراکه این صفت ها تاثیر بیشتری را در تصمیم نهایی ما می گذارند. این روند جلو رفتن در درخت تصمیم گیری تا جایی ادامه می یابد که در هر مسیر دیگر صفتی باقی نمانده باشد. در این حالت درخت تصمیم گیری کاملا ساخته شده است و به پایان رسیده است.

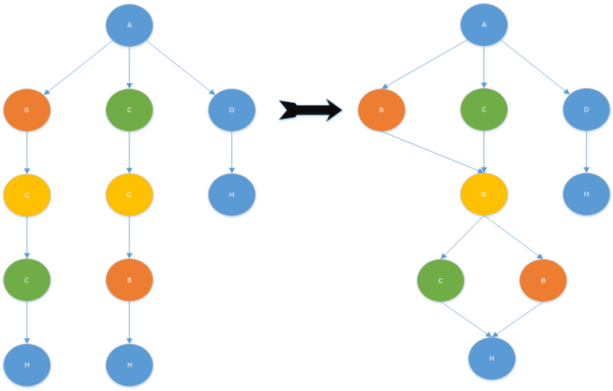
تبدیل درخت تصمیم گیری

برای اینکه بتوان شرط هایی را از این درخت استخراج نمود بصورت If…then…. شود که در آن بتوان از قوانین انجمنی نیز کمک گرفت، ما درخت را تبدیل نمودیم که در موارد بسیاری این درخت تبدیل به گراف می شود و از حالت درخت خارج می شود البته اگر این درخت از نظر ظاهری تبدیل به گراف شود برای ما همچنان بصورت درخت در نظر گرفته می شود بنابراین می توان گفت چیزی مابین گراف و درخت بدست می آید.
در این قسمت ما در هر سطح گره های همنام را با هم ادغام می کنیم و فرزندان آن ها به این گره ادغام شده افزوده می شود. می توان در شکل (۶) این کار را مشاهده نمود تا بهتر بتوان آن را درک نمود.



شکل (۶): مثالی از تبدیل درخت تصمیم گیری

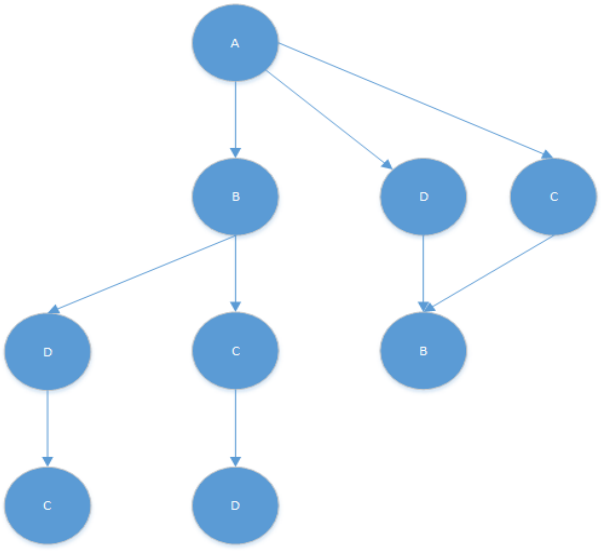
همانطور که در شکل (۶) می توان مشاهده نمود گره های B و C دارای فرزند مشترک G بودند که این دو گره G در یک سطح نیز قرار داشتند. در این حالت این دو گره تبدیل به یک گره شده و فرزندان آن ها نیز به گره جدید افزوده می شود. ممکن است شرایط مانند شکل (۷) پیش آید که در آن دو گره G باهم ادغام شدند ولی حالتی است که در آن گره G اول دارای فرزند C و گره G دوم دارای فرزند B می باشد و این گره ها هر کدام برای گره G مقابل در پیشینیان آن بازدید شده است. در این حالت نیز مشکلی پیش نمی آید و تنها می توان گفت در هنگام نوشتن شرط می توان در نظر گرفت که با داشتن شرط B و C، همراه با And منطقی آن با G آن مسیر پیموده می شود یعنی A And (B OR C) AND G.



شکل (۷): مثالی از حالت خاص تبدیل

در اینجا می رسیم به این قسمت که هر کدام از این صفت ها در هر یک از داده ها می تواند داری مقدار متفاوتی باشد برای مثال در شکل (۸)، G می تواند دارای دو مقدار true و یا false باشد برای این قسمت ما داده ای را انتخاب می کنیم که فراوانی آن در صفت مربوطه در کل داده ها بیشتر است. یعنی برای مثال فراوانی true بیشتر است یا false در بین صفت داده های از نوع G آنگاه آن را بر میگزینیم. در صورتی که فراوانی دو مقدار برابر باشد آنگاه در تصمیم، or را برای این دو مقدار قید میکنیم برای مثال G=true or G=false.

استخراج تصمیمات از گراف تصمیم گیری که در اینجا ساخته شده است. برای اینکار از گره ریشه شروع کرده و به سمت هر یک از برگ ها که برویم یک تصمیم شکل می گیرد.



شکل (۸): مثالی از گراف تصمیم روش پیشنهادی

ماشین بردار پشتبان در مدل پیشنهادی

در این مرحله در ابتدای کار نرمال سازی انجام می شود و سپس نتایج استخراج شده از نرمال سازی، قسمتی از این داده ها به عنوان داده های آموزش استفاده شده و مدل ماشین بردار ایجاد می شود و سپس با استفاده از داده های تست وزن این الگوریتم محاسبه می شود تا اینکه بتوان در ادامه میزان تاثیرگذاری این قسمت از الگوریتم را محاسبه نمود.

شبکه بیزین در مدل پیشنهادی

در این مرحله نیز در ابتدا نرمال سازی انجام می شود و سپس نتایج استخراج شده از نرمال سازی، قسمتی از این داده ها به عنوان داده های آموزش استفاده شده و مدل شبکه بیزین ایجاد می شود و سپس با استفاده از داده های تست وزن این الگوریتم محاسبه می شود تا اینکه بتوان در ادامه میزان تاثیرگذاری این قسمت از الگوریتم را محاسبه نمود.

ارزیابی روش پیشنهادی

در این بخش روش پیشنهادی که در قسمت قبل بیان شد مورد بررسی قرار می گیرد و روش ارائه شده با الگوریتم های معروف به نام ID3، الگوریتم SVM و شبکه بیزین مورد مقایسه قرار می گیرد. همانطور که در قسمت قبل بیان شد راهکار ارائه شده مبتنی بر ID3 است که نسبت به روش ID3 دارای مزایای بیشتری می باشد.

داده ها در ابتدا توسط برنامه به فرمت مناسب برای تحلیل قرار می گیرد یا به عبارتی پیش پردازش ابتدایی صورت می گیرد فایلی با فرمت ARFF ایجاد می باشد که ساختاری مناسب و استاندارد برای تحلیل می باشد.

پیاده سازی در برنامه Visual Studio 2017 و با زبان برنامه نویسی C# صورت گرفته است و در حین کار از کتابخانه هایی weka و zedgraph کمک گرفته شده است.

سیستم مورد استفاده در اینجا دارای سیستم عامل Windows 10 ، دارای ۶ گیگ RAM و Corei7 می باشد.

در این پیاده سازی از دیتاست Quandl (سایت Quandle، ۲۰۱۸) استفاده شده است. Quandl مجموعه کاملی از دیتاست های مربوط به بازار بورس جهانی را دارد که در اینجا از مجموعه ای از آن با نام Tata Global Beverages استفاده شده است. داده های استفاده شده مربوط به سال ۲۰۱۸ می باشند. این مجموعه شامل مجموعه اطلاعاتی می باشد که در جدول (۱) مشخص شده است.

جدول (۱): مشخصات مجموعه داده Tata Global Beverages

Date	Open	High	Low	Last	Close	Total Trade Quantity	Turnover(Lacs)
2018-10-08	208.00	222.25	206.85	216.00	215.15	4642146.0	10062.83
2018-10-05	217.00	218.60	205.90	210.25	209.20	3519515.0	7407.06
2018-10-04	223.50	227.80	216.15	217.25	218.20	1728786.0	3815.79
2018-10-03	230.00	237.50	225.75	226.45	227.60	1708590.0	3960.27
2018-10-01	234.55	234.60	221.05	230.30	230.90	1534749.0	3486.05

ستون های open و close نشان دهنده قیمت پایه و نهایی است که سهام آن در یک روز خاص معامله می شود. high، low و last نشان دهنده حداکثر، حداقل و آخرین قیمت سهم در آن روز است.
Total trade quantity، تعداد سهام خریداری شده یا فروخته شده در روز است و turnover، گردش مالی شرکت خاص در یک تاریخ معین است.
یکی از مواردی که باید در اینجا به آن توجه کرد این است که در این دیتاست بعضی از مقادر موجود نیستند زیرا روزهایی که تعطیل رسمی است مثل آخر هفته و یا تعطیلات دیگر، بازار بورس نیز تعطیل می باشد.

محاسبه سود یا ضرر معمولا به وسیله قیمت close یک سهام برای روز مشخص می شود، بنابراین در این تحقیق ستون close به ستون هدف می باشد و دارای اهمیت بالایی می باشد.

در این تحقیق از نوع دیتاست نظارت شده استفاده شده است یعنی برای هر سهم در نظر گرفته شده است که در آینده شامل رشد خواهد بود و یا نه که در این حالت سهم مورد می بایست انتخاب شود و یا نه. برای اینکه بتوان دیتاستی نظارت شده ایجاد نمود در این تحقیق چنین عمل شده است که برای روند رشد قیمت تنها اگر سهمی دارای رشد قیمت در آینده می باشد به عنوان مناسب و در صورتی که شامل رشد نباشد به عنوان نا مناسب در نظر گرفته شده است. در این تحقیق در نظر گرفته شده است که در ابتدا کار، کاربر دارای سهامی نمی باشد از این جهت نزول قیمت در این روش در نظر گرفته نشده است. یعنی سهامی که ممکن است شامل نزول و کاهش قیمت شوند در نظر گرفته نشده اند. از اینرو در این تحقیق دیتاستی نظارت شده ایجاد و با استفاده از آن سعی شده است تا بررسی ها صورت گیرد.

در این تحقیق الگوریتم پیشنهادی با الگوریتم های ID3، SVM و بیزین مورد مقایسه قرار گرفته است. الگوریتم های ID3 و SVM الگوریتم های پایه روش پیشنهادی می باشند که این روش پیشنهادی از ترکیب این دو روش ایجاد شده است و همچنین روش با یکی از الگوریتم های معروف به نام شبکه بیزین نیز مورد مقایسه قرار گرفته است که در ادامه می توان نتایج پیاده سازی را برای این الگوریتم ها مشاهده نمود.

```
--------Naive Baysian----------
confusionMatrix:
[0,0] = 18 [0,1]=11
[1,0] = 27 [1,1]=124
Correct Prediction Percent = 78.8888888888889%
InCorrect  Prediction Percent = 21.1111111111111%
MeanAbsoluteError(MAE) = 0.215093567412967
MeanSquaredError(MSE) = 0.450450314025597
RelativeAbsoluteError(REA) = 57.1003786873271
Correct Prediction Number = 142
InCorrect Prediction Number = 38
TP: 124
FP: 11
FN: 27
TN: 18
```

شکل (۹): خروجی مربوط به الگوریتم بیزین

```
--------MyAlgorithm----------
confusionMatrix:
[0,0] = 18 [0,1]=9
[1,0] = 27 [1,1]=126
Correct Prediction Percent = 80%
InCorrect  Prediction Percent = 20%
MeanAbsoluteError(MAE) = 0.251683833684513
MeanSquaredError(MSE) = 0.390570453145535
RelativeAbsoluteError(REA) = 66.813909805457
Correct Prediction Number = 144
InCorrect Prediction Number = 36
TP: 126
FP: 9
FN: 27
TN: 18
```

شکل (۱۰): خروجی مربوط به الگوریتم پیشنهادی

```
--------ID3----------
confusionMatrix:
[0,0] = 7 [0,1]=7
[1,0] = 38 [1,1]=128
Correct Prediction Percent = 75%
InCorrect  Prediction Percent = 25%
MeanAbsoluteError(MAE) = 0.343346480854847
MeanSquaredError(MSE) = 0.430065588743028
RelativeAbsoluteError(REA) = 91.147375133409
Correct Prediction Number = 135
InCorrect Prediction Number = 45
TP: 128
FP: 7
FN: 38
TN: 7
```

شکل (۱۱): خروجی مربوط به الگوریتم ID3

```
--------SVM----------
confusionMatrix:
[0,0] = 2 [0,1]=4
[1,0] = 43 [1,1]=131
Correct Prediction Percent = 73.8888888888889%
InCorrect  Prediction Percent = 26.1111111111111%
MeanAbsoluteError(MAE) = 0.261111111111111
MeanSquaredError(MSE) = 0.510990323891863
RelativeAbsoluteError(REA) = 69.3165467625899
Correct Prediction Number = 133
InCorrect Prediction Number = 47
TP: 131
FP: 4
FN: 43
TN: 2
```

شکل (۱۲): خروجی مربوط به الگوریتم SVM

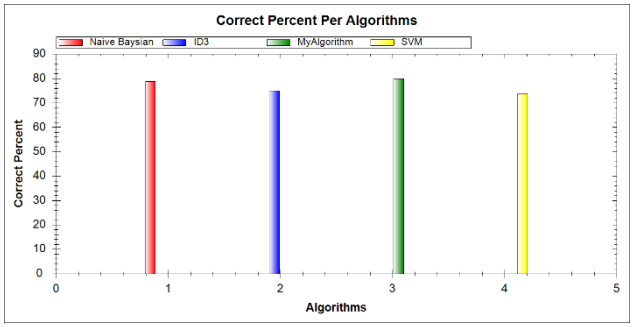
با توجه به نتایج دریافتی می توان به وضوح مشاهده نمود که الگوریتم پیشنهادی با ۸۰ ٪ دقت دارای بالاترین دقت صحت و با ۲۰ ٪ اشتباه دارای کمترین میزان اشتباه می باشد. در جدول شماره (۲) به مقایسه الگوریتم پیشنهادی و سایر الگوریتم‌ها پرداخته می‌شود. کاملا مشخص است که الگوریتم پیشنهادی از سایر الگوریتم‌ها بهتر عمل می کند.

	درصد صحت پیش بینی ها	درصد خطای پیش بینی ها
الگوریتم پیشنهادی	۸۰٪	۲۰٪
الگوریتم بیزین	۷۸.۸۸٪	۲۱.۱۱٪
الگوریتم ID3	۷۵٪	۲۵٪
الگوریتم SVM	۷۳.۸۸٪	۲۶.۱۱٪

جدول (۲): نسبت درصد صحت پیش بینی ها و خطای پیش بینی ها

با توجه به این دقت های بدست آمده کاملا قابل مشاهده می باشد که روش پیشنهادی از دیگر روش ها بسیار بهتر عمل کرده است و در این بین الگوریتم بیزین نیز از الگوریتم های ID3 و SVM نیز بهتر عمل کرده است. همچنین الگوریتم ID3 نیز از الگوریتم SVM بهتر عمل کرده است. در اینجا می توان کاملا مشاهده کرد که در روش ترکیبی پیشنهادی بیان شده می توان با ادغام روش های ID3، شبکه بیزین و SVM که هر یک دارای دقت پایین تری از شبکه بیزین می باشند به دقت بالاتری از شبکه بیزین دست یافت البته باید ذکر کرد که در روش پیشنهادی از ID3 بهبود یافته استفاده شده است که روند بهبود نیز در قسمت قبل بیان شد. این درخت دارای کمترین ارتفاع ممکن می باشد و از اینرو دارای سربار پایین تری می باشد.

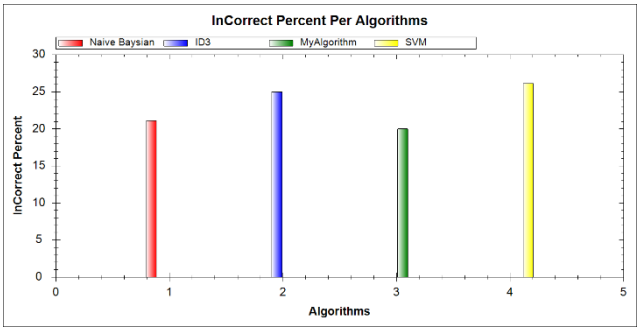
در ادامه نمودارهای بدست آمده از برنامه مورد بررسی قرار گرفته است. اولین نمودار، نمودار مربوط به درصد پیش بینی صحیح در میان داده های آزمون می باشد که می توان در شکل (۱۳) مشاهده نمود.



شکل (۱۳): درصد پیش بینی صحیح در میان داده های آزمون برای الگوریتم پیشنهادی و دیگر الگوریتم های مورد بررسی

همانطور که می توان از این نمودار دریافت روش پیشنهادی ما دارای دقت پیش بینی صحیح بیشتری نسبت به دیگر الگوریتم های مورد بررسی یعنی شبکه بیزین، الگوریتم SVM و الگوریتم ID3 می باشد. این بدین دلیل می باشد که ما در روش پیش بینی خود تنها مواردی از داده های آزمون را در نظر گرفتیم که تاثیر بیشتری را در نتیجه خروجی داشتند و در نتیجه داده هایی که در خروجی تاثیر نداشتند را استفاده نکردیم و بدین شکل زمان تحلیل را بسیار کاهش دادیم و حال انکه الگوریتم های دیگر به دلیل استفاده از تمامی پارامترها دارای دقت کمتری هستند زیرا ممکن است بعضی از پارامتر ها دارای مقادیر دوری باشند که ممکن است هیچ تاثیر در نتیجه خروجی نداشته باشند ولی چون در الگوریتم های دیگر در ساخت مدل برای پیش بینی این پارامترها مورد استفاده قرار گرفته اند باعث ایجاد نویز و کاهش دقت می شوند و در روش پیشنهادی ما چون این پارامترها بی فایده وجود نداشتند در نتیجه دقت در روش پیشنهادی ما افزایش یافت و از دیگر الگوریتم ها بهتر عمل نموده است. همچنین در الگوریتم پیشنهادی از یک روش ترکیبی استفاده شده است که می توان مشاهده نمود به خوبی عمل نموده است و از دیگر روش ها بسیار بهتر عمل کرده است و از روش هایی که بر پایه آن ها ایجاد است نیز بهتر عمل کرده است.

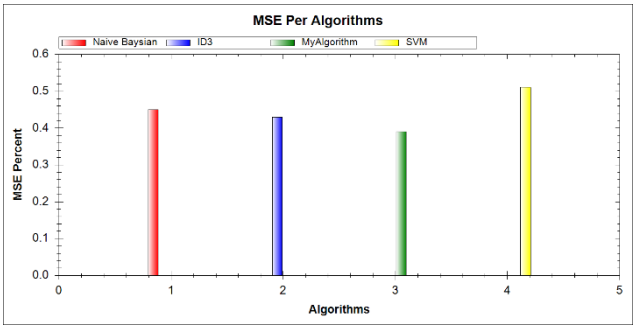
در نمودار ارائه شده در شکل (۱۴) می توان درصد پیش بینی ناصحیح را مشاهده نمود. با توجه به این نمودار می توان درک نمود که روش پیشنهادی به همان دلیل که پیش تر در رابطه با درصد پیش بینی درست گفته شد از باقی روش ها دارای مقدار کمتری است یعنی پیش بینی اشتباه کمتری دارد بنابراین روش پیشنهادی بهتر از باقی روش ها عمل کرده است.



شکل (۱۴): درصد پیش بینی ناصحیح در میان داده های آزمون برای الگوریتم پیشنهادی و دیگر الگوریتم های مورد بررسی

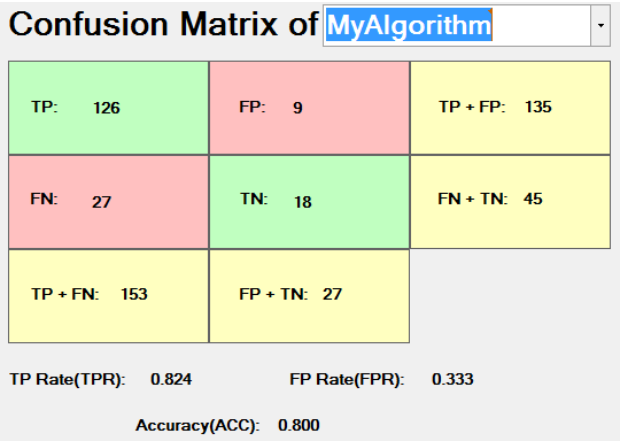
با توجه به نمودار شکل (۱۴) می توان به این قضیه پی برد که خطای پیش بینی تقریبا در الگوریتم های مورد بررسی برابر و نزدیک است ولی الگوریتم پیشنهادی کمتر می باشد زیرا هر چه نرخ صحت بیشتر باشد قاعدتا نرخ غلط پایین تر خواهد بود و این نشان دهنده عملکرد مناسب روش پیشنهادی می باشد. در ادامه می توان مشاهده نمود که نرخ خطای میانگین(MSE) در روش پیشنهادی از تمامی روش های دیگر کمتر می باشد و الگوریتم ID3

نیز از الگوریتم SVM کمتر می باشد و در عین حال الگوریتم بیزین از الگوریتم SVM دارای نرخ خطای کمتری می باشد. این نرخ خطا تنها غلط بودن را بررسی نمی کند بلکه میزان دور بودن جواب پیش بینی شده نسبت به جواب واقعی را نیز محاسبه می کند که در این حالت مشاهده می شود که الگوریتم پیشنهادی بسیار بهتر از دیگر الگوریتم ها رفتار می کند. اگر در این نمودار توجه شود می توان به این قضیه پی برد چرا که الگوریتم ID3 دارای پیش بینی غلط بیشتری از الگوریتم بیزین است ولی چون در آنجا میزان دور بودن در نظر گرفته نمی شد ID3 بدتر از شبکه بیزین بود ولی می توان در شکل (۱۵) مشاهده نمود که الگوریتم ID3 دارای MSE کمتری از الگوریتم بیزین می باشد که این یعنی دارای نرخ خطای کمتری می باشد. به طور کلی در علم داده کاوی MSE دارای اهمیت بسیار بالایی می باشد و دارای اعتبار بسیار زیادی است.

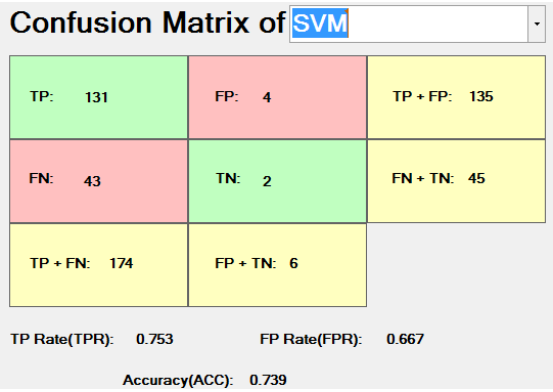


شکل (۱۵): میزان معیار MSE در میان الگوریتم پیشنهادی و دیگر الگوریتم های مشابه مورد بررسی

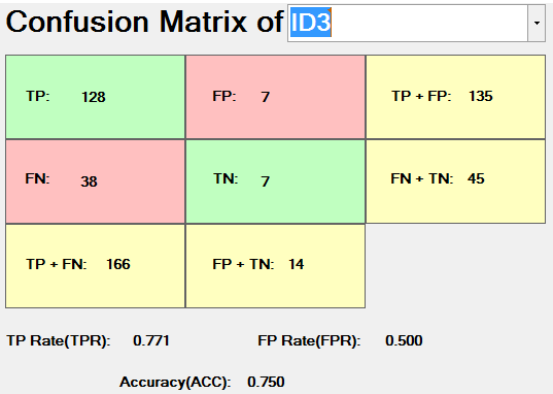
در ادامه Confusion Matrix مربوط به روش پیشنهادی و دیگر روش های مورد بررسی در این تحقیق نشان داده شده است.



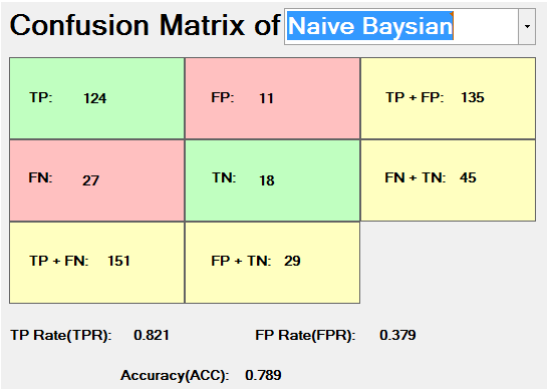
شکل (۱۶): جدول Confusion matrix روش پیشنهادی



شکل (۱۷): جدول Confusion matrix روش SVM



شکل (۱۸): جدول Confusion matrix روش ID3



شکل (۱۹): جدول Confusion matrix روش بیزین

در اینجا می توان مشاهده نمود که روش پیشنهادی دارای دقت بالاتری می باشد زیرا در این جدول دارای مقدار TP و TN بیشتری از دیگر الگوریتم ها می باشد و در کنار آن نیز دارای FP و FN کمتری از دیگر الگوریتم های مورد بررسی در اینجا می باشد. زیرا هر چه یک الگوریتم دارای TP و TN بیشتری باشد یعنی متقلب بودن و یا نبودن های در مجموعه داده های تست را به مقدار بیشتری درست تشخیص داده است و FP و FN نشان دهنده برعکس این قضیه یعنی پیش بینی اشتباه می باشد.

بحث و نتیجه گیری

با توجه به گسترش بازارهای مالی و توجه زیاد محققین در این زمینه به منظور استفاده از سیستم های قاعده محور به منظور تصمیم گیری های سریع و با ریسک حداقل و به دور از اشتباهات انسانی، طراحی و توسعه و یا بهبود این سیستم ها می تواند مزیت رقابتی بسیار خوبی برای سرمایه گذاران باشد. از این رو، در این تحقیق، یک روش مبتنی بر مجموعه راف و تحلیل سلسله مراتبی برای تشخیص پارامترهای تاثیرگذار و پیش بینی انتخاب مناسب سهام برای سبد بهینه سهام ارائه شده است. در پژوهش ارائه شده در این تحقیق تعداد ویژگی ها کاهش یافته است تا سربار محاسباتی کاهش یابد و همچنین در مواردی استفاده از ویژگی هایی که تاثیری در خروجی ندارند و ممکن است باعث ایجاد نویز شود، از الگوریتم مجموعه راف و تحلیل سلسله مراتبی در جهت تشخیص ویژگی های موثر استفاده شده است و تنها ویژگی هایی انتخاب شده اند که دارای بیشترین تاثیرگذاری در خروجی باشند و بتوان نتایج دقیق تر را با سربار محاسباتی پایین تری دست پیدا کرد. از طرف دیگری در این تحقیق از درخت تصمیمی استفاده شده است که id3 بهبود داده شده است و این بهبود باعث کاهش گره های درخت و همچنین کاهش تعداد لایه های درخت و در نتیجه کاهش به شدت زیاد سربار محاسباتی و افزایش دقت محاسبات می شود. در این پژوهش راهکاری برای ارزیابی و پیش بینی بینی سهام ارائه شد و مشاهده شد که روش ارائه شده در اینجا دارای عملکرد مناسبی بوده و بهبود نسبتا بالایی را نسبت به الگوریتم های پایه خود یعنی ID3، شبکه بیزین و SVM نشان داده است که روش پیشنهادی نسبت به این الگوریتم ها دارای بهبود عملکرد می باشد. با توجه به نتایج دریافتی مدل ارائه شده با ۸۰ ٪ دقت دارای بالاترین دقت صحت و با ۲۰ ٪ اشتباه دارای کمترین میزان اشتباه می باشد.

ارزیابی عملکرد پیش بینی قیمت سهام و انتخاب سبد بهینه با توجه به تغییرات قیمت آینده سهام با استفاده از روش ارائه شده در این تحقیق نشان می دهد که این روش قادر به پیش بینی نوسانات شدید با الگوهای غیرخطی می باشد که بیان گر این است که پیش بینی ها رضایت بخش است. هم چنین، بررسی عملکرد پیش بینی قیمت سهام بدون استفاده از روش پیشنهادی این تحقیق نشان می دهد که هر چند این روش قادر است که برخی از نوسانات رخ داده در قیمت سهام را پیش بینی نماید. همچنین نشان داده شد که روش پیشنهادی می تواند در فراز و نشیب های مختلف نیز نتایج بسیار خوبی را نشان دهد و دارای دقت بالایی باشد و از معایب روش پیشنهادی می توان به سربار زمانی اشاره کرد که کمی فرایند تحلیل را زمانبر می کند البته اگر در این حالت از ماشین های قوی استفاده شود روند می تواند سریعتر نیز اجرا شود.

پیش بینی دقیق تغییرات شدید قیمت در مقایسه با تغییرات با دامنه و شدت کمتر که بالطبع، به سود و زیان کمتری نیز منجر می شود، بسیار حائز اهمیت تر می باشد.

ارزیابی کارایی رویکرد پیشنهادی برای پیش بینی روند تغییرات قیمت سهام از طریق محاسبه شاخصRMSE نشان می دهد که بهترین نتایج مربوط به روش پیشنهادی می باشد. با استفاده از رویکرد ارائه شده در این تحقیق می توان بهترین سبد سهام را انتخاب نمود زیرا سهامی در سبد بهینه سهام قرار می گیرد که دارای بیشترین تغییرات قیمت در جهت افزایش برای آینده باشد و در این حالت می توان به نتایج بسیار قابل قبول دست یافت و در نتیجه به بهترین روش برای انتخاب سبد بهینه دست یافت.

البته از نتیجه روش بیان شده در این تحقیق می توان علاوه بر انتخاب سبد بهینه سهام، برای فروش سهامی که احتمال سقوط آن ها در آینده پیش بینی می شود نیز استفاده نمود بنابراین می توان از نتیجه این تحقیق نه تنها برای انتخاب سبد بهینه سهام برای خرید بلکه برای فروش سهام بی ارزش و محتمل بر ضرر نیز استفاده نمود.

پیشنهادهایی برای تحقیقات آتی

۱) در این روش پیشنهادی از آنتروپی استفاده شده است ولی می توان از روش های دیگری نیز استفاده نمود و یا این روش را با روش های دیگری ادغام نمود. برای مثال در صورتی که این روش را با روشی مانند Gain که تابع ارزش می باشد ادغام نماییم، به احتمال زیاد دارای عملکرد بهتری می باشد چراکه آنتروپی نیز دارای معایبی می باشد ولی سرعت آن بالاست و در اینجا هم ما به دنبال روشی بودیم که دارای سرعت بالا باشد ولی می توان با ادغام این روش و یا روش های جایگزین دقت این روش ارائه شده را به شدت افزایش داد.

۲) می توان روش پیشنهادی را با بهبود در الگوریتم C4.5 نیز استفاده نمود یعنی روش پیشنهادی را روی C4.5 با بهبودی مشابه بهبود که روی ID3 در این تحقیق انجام شد، انجام داد تا عملکرد آن افزایش یابد البته نمی توان به طور قطع گفت که عملکرد آن بهتر می شود بلکه می بایست این روش مورد آزمایش قرار گیرد تا صحت عملکرد بررسی شود.

فهرست منابع

* تهرانی، رضا؛ هندیجانی زاده، محمد و نوروزیان لکوان، عیسی(۱۳۹۴)، ارائه رویکردی جدید برای مدیریت فعال پرتفوی و انجام معاملات هوشمند سهام با تاکید بر نگرش انتخاب ویژگی، فصلنامه علمی پژوهشی دانش سرمایه گذاری، سال چهارم، شماره سیزدهم، بهار ۱۳۹۴.
* راعی، رضا؛ پویان فر، احمد. (۱۳۹۸). مدیریت سرمایه گذاری پیشرفته، تهران، انتشارات سمت.
* رستمی، ستیلا(۱۳۹۵)، بررسی کارایی الگوریتم یادگیری ماشین در پیش بینی شاخص بورس و اوراق بهادار تهران، منتشر شده در دومین کنفرانس ملی علوم مدیریت نوین و برنامه ریزی فرهنگی اجتماعی ایران در سال ۱۳۹۵.

* شریعت پناهی، سیدمجید، جعفری، ابوالفضل. ۱۳۹۳، مدیریت سرمایه گذاری، تهران، انتشارات اتحاد.
* شفیعی، حامد، یک مدل فازی شبکه عصبی برای پیش بینی قیمت سهام، پایان نامه مقطع کارشناسی ارشد، دانشکده صنایع، دانشگاه صنعتی شریف، ۱۳۹۴.
* عالم تبریز، اکبر؛ محمدعلی افشاری، محمدحسن ملکی و جواد محمدی. ۱۳۸۹؛ انتخاب بهینه سبد سهام با استفاده از مدل شبکه عصبی- مصنوعی، اریما و مدل مارکوئیتز در بورس اوراق بهادار تهران، اولین کنفرانس بین المللی مدیریت و نوآوری، شیراز.
* میرعلوی، سیدحسین؛ پرزمانی، زهرا. مدلی جهت پیش بینی قیمت سهام با استفاده از روش های فرا ابتکاری و شبکه های عصبی، فصلنامه مهندسی مالی و مدیریت اوراق بهادار، ۱۳۹۸، شماره ۴۰.
* خنده خوش، مژگان؛ حقیقی نیت،رضا(۱۳۹۵)، در نظر گرفتن عوامل مؤثر در پیش‌بینی شاخص قیمت بورس تهران با بهبود الگوریتم بهینه‌سازی ملخ در انتخاب بهترین نمونه‌ها در مدل آموزش چندتایی شبکه عصبی، سومین کنفرانس بین المللی مدیریت، اقتصاد و حسابداری.
* https://www.quandl.com
* Site: https://virgool.io/@sanaye20.ir, (2020).
* Zhang Qinghua, Qin Xie, GuoyinWang, “A survey on rough set theory and its applications”, CAAI Transactions on Intelligence Technology, Volume 1, Issue 4, October 2016, Pages 323-333.
* L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
* Site: https://en.wikipedia.org/wiki/Decision_tree_learning, (2020).
* Site: https://en.wikipedia.org/wiki/Predictive_analytics, (2019).
* Bertsimas, D., Shioda, R. (2009). Algorithm for Cardinality constrained quadratic optimizatio, Computational Optimization and Applications, Vol. 43, pp. .۱22.
* Chen, Yan, et al. "A genetic network programming with learning approach for enhanced stock trading model." Expert Systems with Applications 36.10: 12537-12546, 2015.
* J.R. Quinlan. C4. 5: programs for machine learning. Morgan Kaufmann, 1993.
* T. Hastie, R. Tibshirani and J. Friedman. Elements of Statistical Learning, Springer, 2009.
* Chen, Yan, and Xuancheng Wang. "A hybrid stock trading system using genetic network programming and mean conditional value-at-risk." European Journal of Operational Research 240.3: 861-871, 2015.
* Chou, Yao-Hsin, et al. "Intelligent stock trading system based on qts algorithm in japan's stock market." Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013.
* Easton, Peter D. "PE ratios, PEG ratios, and estimating the implied expected rate of return on equity capital." The accounting review 79.1: 73-95, 2016.
* Glasserman, Paul. Monte Carlo methods in financial engineering. Vol. 53. Springer Science & Business Media, 2013.
* M. Yoda, Predicting the Tokyo Stock Market. In: Trading on The Edge, Deboeck, G.J. (Ed.), John Wiley & Sons Inc., 2016, pp. 66-79.

* Yang, Yang, et al. "GNP-Sarsa with subroutines for trading rules on stock markets." Systems Man and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013.
* Zahedi, Javad, and Mohammad Mahdi Rounaghi. "Application of artificial neural network models and principal component analysis method in predicting stock prices on Tehran Stock Exchange." Physica A: Statistical Mechanics and its Applications 438: 178-187, 2015.
* Site: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/, (2020).
* Site: https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e, (2020).
* Site: https://www.bayesserver.com/docs/introduction/bayesian-networks, (2020).

Financial Economics
Vol. (12) Issue (48) Winter 2024

Optimal Portfolio Selection using Machine Learning Algorithms

Mohammad Bagher Yazdani Khodashahri
Ph.D Student, Department Of Accounting, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran.

Seyed Hossein Nasl Mosavi
Department Of Accounting, Qaemshahr Branch, Islamic Azad University, Qaemshahr, Iran.

Mirsaeid Hosseini Shirvani
Department Of Computer, Sari Branch, Islamic Azad University, Sari, Iran.

Abstract
Choosing the right portfolio is always one of the most important issues for investors. The price trend is predicted using technical analysis or basic analysis. Technical analysis focuses on market performance, while the focus of fundamental analysis is on the mechanism of supply and demand, and these changes prices. The existence of a solution to predict growth or decrease in stocks has been studied as a basic need in this study. In the present study, with the help of a monitoring dataset, a solution based on Raff collection algorithms and hierarchical analysis to reduce the feature and decision tree algorithms, backup vector machine, and business network have been used for prediction. This proposed solution has been implemented using language and compared with different solutions, and the research results have shown that the proposed method with 80% accuracy of prediction and 20 errors in prediction has the highest accuracy and the lowest error rate among the methods compared
Keywords: Support vector machines, Bayesian network, improved decision tree, Rough set, portfolio selection