

شناسایی و حذف داده‌های پرت و تکراری در شبکه‌های حسگر بی‌سیم با استفاده از تکنیک‌های طبقه‌بندی (مطالعه موردی: داده‌های هواشناسی کاشان)

مهدی اسماعیلی^{۱*}، عظیمه شریف^۲

۱- استادیار، گروه کامپیوتر و فناوری اطلاعات، واحد کاشان، دانشگاه آزاد اسلامی، کاشان، ایران

۲- کارشناسی ارشد، گروه کامپیوتر و فناوری اطلاعات، واحد کاشان، دانشگاه آزاد اسلامی، کاشان، ایران

چکیده

شبکه‌های حسگر بی‌سیم نسل جدیدی از شبکه‌ها هستند که در واقع از تعداد زیادی گره‌های حسگر پراکنده در محیط تشکیل شده‌اند و هر کدام به‌طور خودمختار و با همکاری سایر گره‌ها هدف خاصی را دنبال می‌کنند. از آنجایی که انتقال داده به‌عنوان یکی از مهم‌ترین عملیات مصرف انرژی در شبکه‌های حسگر بی‌سیم می‌باشد، بنابراین کاهش میزان انتقال داده در این شبکه‌ها منجر به کاهش مصرف انرژی و در نتیجه طول عمر بیشتر این شبکه‌ها می‌گردد. روش‌هایی که برای کاهش میزان انتقال داده در شبکه‌های حسگر بی‌سیم باهدف ذخیره انرژی، اغلب استفاده می‌شود شامل تجمع داده‌ها، پالایش و پیش‌بینی داده‌ها می‌باشد. از سوی دیگر، استفاده از پروتکل‌های مسیریابی مناسب می‌تواند در حفظ انرژی مصرفی گره‌ها نقش مؤثری داشته‌باشند. در این پژوهش سعی شده از طریق شناسایی داده‌های پرت توسط برخی از تکنیک‌های طبقه‌بندی و پروتکل‌های مسیریابی سلسله‌مراتبی، انرژی مصرفی در کل شبکه کاهش یابد. این آزمایشات روی داده‌های حسگرهای ایستگاه خودکار هواشناسی کاشان انجام شد. نتایج نشان می‌دهد که با به‌کارگرفتن تکنیک ماشین بردار پشتیبان با کرنل چندجمله‌ای از درجه ۳ و با ضریب تنظیم‌گر ۱۰ در شبکه‌ی حسگر پیشنهادی می‌توان حدود ۹۲٪ داده‌ها را در مدت ۲ ثانیه به‌درستی شناسایی نموده و بنابراین با فیلتر کردن داده‌های پرت و داده‌های تکراری و ارسال داده جدید و صحیح، به صرفه‌جویی انرژی و در نتیجه افزایش طول عمر شبکه کمک نمود.

کلید واژه‌ها: شناسایی داده‌پرت، شبکه حسگر بی‌سیم، تکنیک‌های طبقه‌بندی، داده‌کاوی، کاهش مصرف انرژی

مقدمه

شبکه‌های حسگر بی‌سیم متشکل از تعداد زیادی گره‌های حسگر می‌باشند که به صورت تصادفی در محیط پراکنده شده‌اند. بنابراین، مکان قرارگیری این گره‌ها از قبل تعیین شده و مشخص نیست که باعث می‌شود بتوانیم آن‌ها را در مکان‌های خطرناک یا غیرقابل دسترس رها کنیم. از سوی دیگر، هریک از گره‌های حسگر به صورت خودمختار عمل می‌کنند و همچنین توانایی همکاری با سایر گره‌های حسگر را به منظور دستیابی به هدف دارند. در شبکه‌های حسگر بی‌سیم مسئله انرژی و استفاده بهینه از آن یکی از مهم‌ترین چالش‌ها در برخورد با چنین شبکه‌ها است، زیرا که گره‌های حسگر بی‌سیم معمولاً از باتری تغذیه می‌کنند و طول عمر شبکه به طول عمر باتری بستگی دارد. روش‌های مختلفی به منظور صرفه‌جویی در مصرف انرژی برای شبکه‌های حسگر بی‌سیم پیشنهاد شده‌اند که مهم‌ترین آن‌ها عبارتند از: کاهش میزان انتقال داده، استفاده از پروتکل‌های مسیریابی و خوشه‌بندی در شبکه‌های حسگر بی‌سیم. راه‌حل و روش مناسب برای برخورد با مسئله انرژی کاملاً به نوع شبکه و کاربرد آن بستگی دارد.

انتقال داده در شبکه‌های حسگر بی‌سیم به‌عنوان یکی از مهم‌ترین عملیات مصرف انرژی شناخته شده است (Zu et al. 2012) هم‌چنین طول عمر باتری با میزان انتقال داده‌ها در یک گره حسگر به‌طور قابل ملاحظه‌ای کاهش می‌یابد. یکی از راه‌های کاهش میزان انتقال داده، شناسایی داده‌های پرت است تا بتوان از انتقال آن‌ها در شبکه جلوگیری نمود.

در شبکه‌های حسگر بی‌سیم، داده‌هایی که از الگوهای عادی منحرف شده‌اند به‌عنوان پرت شناخته می‌شوند و از آن جهت اهمیت دارند که ممکن است بیانگر ناهنجاری مهم و قابل‌توجهی باشند. پس از شناسایی داده‌های پرت بسته به تصمیم سیستم یا می‌توان آن‌ها را حذف کرده و یا در حافظه‌ی حسگر به‌منظور بررسی‌های بیش‌تر ذخیره نمود و به‌طور کلی از ادامه‌ی انتقال آن‌ها در شبکه خودداری کرد که منجر به کاهش مصرف انرژی در کل شبکه خواهد شد. از سوی دیگر، در صورتی که فاصله گره‌های حسگر نسبت به یکدیگر کم باشد هرگاه یک رویداد در ناحیه‌ای رخ دهد،

ممکن است توسط چندین گره حس شود و همه‌ی آن‌ها یک پیغام را ارسال کنند. همچنین در بسیاری از کاربردهای نظارتی شبکه حسگر بی‌سیم مانند هواشناسی که حسگرها به صورت متوالی عملیات دریافت داده‌ها از محیط را انجام می‌دهند، به دلیل این‌که داده‌های محیطی مانند دمای هوا و سایر پارامترها معمولاً به‌کندی تغییر می‌کنند، ممکن است دریافت‌های متوالی یک گره حسگر مشابه باشند و منجر به ارسال داده‌های تکراری در شبکه شوند که علاوه بر ایجاد ازدحام در شبکه باعث کاهش شدید انرژی می‌شود. از این رو، با جلوگیری از ایجاد داده‌های تکراری یا پالایش آن‌ها نیز می‌توان انرژی مصرفی در کل شبکه را کاهش داد.

هدف این پژوهش، استفاده از تکنیک‌های طبقه‌بندی رایج و به‌دست آوردن مقادیر بهینه برای پارامترهای آن به‌منظور ساخت مدلی جهت شناسایی داده پرت با سرعت و دقت بالا می‌باشد. در این پژوهش سعی شد تا با شناسایی داده‌های پرت و تکراری حس شده توسط حسگرها و جلوگیری از انتقال آن‌ها در شبکه حسگر بی‌سیم که از ستون فقرات مجازی برای ارسال چندگامی استفاده می‌کند، انرژی مصرفی در کل شبکه کاهش یابد.

در بخش دوم پیشینه تحقیق مرور خواهد شد. در بخش سوم روش کار و الگوریتم‌های مورد استفاده بررسی خواهد شد. در بخش چهارم ساختار شبکه حسگر بی‌سیم پیشنهادی مطرح خواهد شد. در بخش پنجم ارزیابی و در بخش ششم نتیجه‌گیری و کارهای آینده ذکر می‌شود.

مروری بر پیشینه تحقیق

اخیراً، بسیاری از شبکه‌های حسگر بی‌سیم در دنیای واقعی به منظور جمع‌آوری داده‌های خام با ارزش استفاده شده‌اند. در تجزیه و تحلیل این داده‌ها یک گام مهم تشخیص داده‌های غیرعادی (شامل پرت، نویز^۱ و...) و حذف آن‌ها است. در (Shen et al. 2013) یک طرح کلی از روش‌های تشخیص داده‌های پرت ارائه شده است.

SVM^۲ یکی از تکنیک‌های یادگیری ماشین است. عملیات جدا کردن داده‌های نرمال از داده‌های غیرعادی با استفاده از

1 Noise

2 Support Vector Machine

اصلی و یک رویکرد تک متغیره مبتنی بر آمار در شبکه‌های حسگر بی‌سیم بررسی شدند.

همچنین در (Bastam, 2008) الگوریتمی ارائه شده که می‌تواند بر محیط به‌طور مکرر نظارت کرده، تغییرات ناگهانی در داده‌ها را اطلاع‌رسانی کند و اگر تغییرات به بیش از حد آستانه‌ی داده رسیده، هشدار دهد. در این روش داده‌های حس شده به دودسته‌ی داده‌ی نرمال و داده‌ی اضطرابی تقسیم می‌شود. به این صورت که وقتی داده‌ی α توسط گره‌ی حسگر دریافت شد، مقایسه‌ای بین α و داده‌ی دریافتی قبلی انجام می‌شود و در صورتی که اختلاف بین آن‌ها زیاد نباشد به‌عنوان داده‌ی نرمال در نظر گرفته شده و در حافظه‌ی گره حسگر ذخیره می‌شود، اما اگر اختلاف آن‌ها زیاد باشد به‌عنوان داده‌ی اضطرابی در نظر گرفته شده و بلافاصله به گره مرکزی منتقل می‌شود.

با توجه به این‌که در بسیاری از کاربردهای نظارتی شبکه حسگر بی‌سیم مثل سیستم جمع‌آوری اطلاعات محیطی برای خانه‌های هوشمند، داده‌های محیطی مثل دما یا روشنایی معمولاً بسیار کند با زمان تغییر می‌کنند، بنابراین دریافت‌های متوالی یک گره حسگر ممکن است مشابه باشند که می‌توان مقادیر تکراری را حذف و تنها مقدار اختلاف داده دریافتی از مقادیر قبلی را منتقل نمود. برای دستیابی به این هدف در (Sepasi zahmati, 2006) الگوریتمی مشابه آنچه که در بالا ذکر شد برای تشخیص داده تکراری در شبکه حسگر بی‌سیم و حذف آن پیشنهاد شده است.

در این الگوریتم مقدار Δ به عنوان حد آستانه برای انتقال داده‌ها تعریف می‌شود، در هر گره حسگر پس از دریافت داده x_n ، اختلاف x_n با داده دریافتی قبلی (x_{kp}) محاسبه شده و اگر مقدار حاصل بزرگتر از Δ باشد داده منتقل می‌شود، علاوه بر این، به جای انتقال کل داده فقط مقدار اختلاف δ منتقل می‌شود، بنابراین اندازه بسته می‌تواند کوتاه‌تر شود. همچنین اگر داده‌های حس شده درجه‌ی کمی از تغییر را داشته باشند، نرخ کاهش داده می‌تواند به ۹۰٪ برسد.

در روش دیگری که در (Zu et al. 2012) ارائه شده، از ارزش اطلاعات داده‌های حسگر برای پالایش داده حس شده استفاده می‌شود. ایده‌ی اصلی این است که دیدگاه کاربر برای تعیین ارزش اطلاعات داده‌ی حسگر ثبت شود. با یک چنین

تکنیک SVM در (Fawzy et al. 2013) ارائه شده است.

تکنیک‌های تشخیص داده‌پرت مبتنی بر SVM تک کلاسه از همبستگی^۱ مکانی و زمانی موجود بین داده‌های حسگر برای تشخیص داده پرت استفاده می‌کنند، اما این تکنیک برای نمونه‌های آموزشی^۲ در مقیاس بزرگ سربار زمانی زیادی در پردازش دارد.

در (Gil et al. 2016) تکنیکی مبتنی بر بیز برای کشف داده‌های پرت محلی در جریان داده حسگر تعریف شد. بر اساس آنچه که در (Vy et al. 2012) ذکر شده، ترکیب برخی از روش‌ها نیز برای تشخیص پرت می‌تواند استفاده شود. برای مثال، ترکیبی از روش‌های نزدیک‌ترین همسایگی و SVM برای تشخیص پرت پیشنهاد شده که از روش نزدیک‌ترین همسایگی برای کاهش مقیاس نمونه‌های آموزشی استفاده می‌کند. این روش می‌تواند زمان آموزش را کوتاه‌تر کرده و زمان را بهینه‌کند اما برای داده‌ها در مقیاس بزرگ روش نزدیک‌ترین همسایگی مصرف زمانی و مکانی قابل توجهی دارد.

و یا در (Li, 2014) ترکیبی از روش SVM و نرخ تفکیک فیشر برای شناسایی داده‌های زائد و تکراری پیشنهاد شده که ابتدا خوشه‌بندی توزیع شده روی داده‌های حسگر اجرا می‌شود، سپس از طریق نرخ تفکیک فیشر مرزهایی برای تفاوت قائل شدن بین داده‌ی خوشه‌بندی شده و داده‌ی پراکنده در خوشه‌ها به دست می‌آید. داده‌های خوشه‌بندی شده به‌عنوان داده‌ی زائد در نظر گرفته شده و حذف می‌شوند. بنابراین تعداد نمونه داده‌ها جهت آموزش SVM تا حد زیادی کاهش می‌یابد اما دقت طبقه‌بندی در این روش کمی پایین‌تر از روش آموزش SVM افزایشی سنتی می‌باشد.

در (Esmaeili, 2016) تکنیکی ترکیبی از روش‌های نزدیک‌ترین همسایه و خوشه‌بندی را به‌منظور محاسبه‌ی وقایع جالب برای تشخیص داده‌ی پرت پیشنهاد کردند.

در (Esmaeili, 2013) سه تکنیک تشخیص پرت شامل تکنیک یادگیری ماشین، روش‌های مبتنی بر تحلیل مؤلفه

1 Correlation
2 Training Set

مرحله پیش‌پردازش داده‌ها یکی از مهم‌ترین مراحل در فرآیند داده‌کاوی می‌باشد؛ در این پژوهش مرحله پیش‌پردازش شامل عملیات تبدیل داده‌ها، برچسب‌گذاری داده‌ها و نرمال‌سازی داده‌ها می‌باشد.

از آنجایی که شکل مناسب داده‌ها به‌عنوان ورودی الگوریتم‌های داده‌کاوی نقش بسزایی در این فرآیند بازی می‌کنند و هم‌چنین به‌دلیل این‌که پژوهش روی داده‌های هواشناسی اجرا شده که به‌صورت گذشته ثبت می‌شوند؛ در ابتدا باید با ایجاد تغییراتی شکل داده‌ها به فرمت مناسب برای استفاده در الگوریتم‌های طبقه‌بندی تبدیل گردند. بدین‌منظور عملیاتی مانند تغییر ساعت از نظر زمانی نسبت به محل مورد اندازه‌گیری، دما و فشار و ... اجرا گردید. سپس جهت تشخیص داده‌های پرت و ساخت مدل اولیه باید برچسب‌گذاری داده‌ها اجرا شود که یکی از ساده‌ترین راهکارها برای رسیدن به این هدف استفاده از توابع آماری است. محاسبه میانگین و انحراف استاندارد و تعریف تابعی از این دو مقدار برای سطح آستانه داده‌ها می‌تواند در تشخیص داده‌های خارج از محدوده مفید باشد. این روش برای داده‌هایی که توزیع آن‌ها شناخته شده باشد، مناسب است. برای تشخیص وضعیت توزیع داده‌ها روش‌های متعددی وجود دارد که در شکل ۱ دیده می‌شود.

پالایش کاربر- محوری میزان انتقال داده می‌تواند کاهش یابد درحالی‌که اطلاعات کافی از دیدگاهی که کاربر ارائه می‌دهد، گرفته می‌شود. بنابراین، یک روش پالایش داده‌ی محلی توصیف شده، که دو مفهوم متفاوت پیش‌بینی داده‌ها و ارزش اطلاعات را در مراحل بعدی پالایش به‌کار می‌گیرد.

روش تحقیق و الگوریتم‌های مورد استفاده

در این پژوهش کارایی روش SVM روی داده‌های هواشناسی به‌منظور تشخیص پرت، با روش‌های رگرسیون و ترکیبی از لحاظ معیارهای صحت، حساسیت، میانگین هارمونیک و هم‌چنین سرعت اجرا مورد مقایسه قرار گرفت. به‌منظور بررسی دقیق‌تر، داده‌ها بر اساس مقادیر مختلف K به نمونه‌های آموزشی و آزمایشی دسته‌بندی شدند و سه کرنل ضرب داخلی، شعاعی و چندجمله‌ای از لحاظ قدرت طبقه‌بندی و تشخیص روی این نوع از داده‌ها مورد ارزیابی قرار گرفتند. هدف این مقاله بررسی کارایی روش SVM برای تشخیص پرت روی داده‌های هواشناسی در مقایسه با روش‌های رگرسیون و ترکیبی از لحاظ متریک‌ها و سرعت اجرا در شبکه حسگر بی‌سیم می‌باشد. روند کلی مراحل انجام شده در این پژوهش به صورت زیر است.

پیش‌پردازش داده‌ها

ماهیت روش‌ها	انواع روش‌ها	
	روش‌های عددی	روش‌های گرافیکی
توصیفی	کشیدگی کجی	نمودار ساقه و برگ نمودار جعبه‌ای نمودار هیستوگرام
مبتنی بر نظریه	آزمون کولموگروف-اسمیرنوف آزمون شاپیرو-ویلک	نمودار احتمال-احتمال (P-P) نمودار چارک-چارک (Q-Q)

شکل ۱ - انواع روش‌های سنجش نرمال بودن داده‌ها

باشد (بیشتر از ۳۰ یا ۴۰ مشاهده)، تخطی از فرض نرمال بودن مشکل آن‌چنانی در محاسبات ایجاد نمی‌کند و در صورتی که داده متشکل از صدها مشاهده باشد توزیع داده آنقدرها مهم نیست. (به دلیل قضیه حد مرکزی). اکثر آزمون‌های نرمال بودن، در صورتی که تعداد مشاهدات داده‌ها زیاد باشد حساسیت زیادی خواهند داشت و

این آزمون‌های سنجش نرمال بودن داده‌ها مانند کولموگروف و ... صرفاً فرض صفر پیروی داده از توزیع نرمال را علیه فرض جایگزین نرمال نبودن بررسی می‌کنند و در تصمیم‌گیری این موضوع که آیا باید از روندهای آماری مبنی بر توزیع نرمال استفاده نمود یا نه ارزش خاصی ندارند. در حالت کلی اگر تعداد داده‌ها به اندازه کافی زیاد

کوچک‌ترین انحرافی (به عنوان مثال مقدار بسیار کمی چولگی یا skewness) در هیستوگرام داده‌ها، احتمال رد فرض صفر نرمال بودن را به شدت بالا می‌برد و از طرف دیگر هنگامی که تعداد مشاهدات کم است (زیر ۳۰ یا ۴۰ مشاهده) یک داده با هیستوگرام کاملاً نامتقارن هم ممکن است تست نرمال بودن را با موفقیت پشت سر بگذارد. لذا برای اطمینان از نوع توزیع داده‌ها در این پژوهش، بررسی نرمال بودن از دو روش محاسبه چولگی و آزمون کولموگروف-اسمیرنوف انجام شد. سپس با توجه به نوع داده‌ها که تقریباً توزیع متوازن و مشخصی دارند، از روش "انحراف معیار ± 3 میانگین = حدآستانه" برای تشخیص داده‌های پرت استفاده شد به طوری که داده‌هایی که در محدوده‌ی دومقدار حدآستانه قرار نمی‌گیرند، به عنوان داده-ی پرت شناخته شدند. همچنین جهت تحلیل دقیق‌تر داده‌ها، نرمال‌سازی با استفاده از معادله (۱) زیر انجام شد که در آن Min و Max به ترتیب حداقل و حداکثر مقادیر موجود برای صفت خاصه‌ی مورد نظر می‌باشد و مقدار $NewMin = 0$ و $NewMax = 1$ در نظر گرفته شده است.

$$(1) x = \frac{x - Min}{Max - Min} (NewMax - NewMin) + NewMin$$

تفکیک داده‌ها

در تکنیک‌های طبقه‌بندی روش اساسی درسنجش مدل به-این ترتیب است که ابتدا با کمک بخشی از داده‌ها مدل ساخته شده و با استفاده از نمونه‌های باقی‌مانده این مدل ارزیابی می‌شود. برای افراز داده‌ها به مجموعه داده‌های آموزشی و آزمایشی تکنیک‌های مختلفی وجود دارد که K-fold cross-validation، Random Subsampling، Holdout از آن جمله‌اند.

استفاده از تکنیک‌های طبقه‌بندی در جهت

تشخیص پرت

روش‌های نظارت شده‌ای مثل طبقه‌بندی سعی می‌کنند تا رابطه‌ای میان صفات خاصه ورودی (متغیر مستقل) را با یک یا چندین صفت خاصه هدف (متغیر وابسته) کشف کنند و در نهایت این رابطه با یک ساختار به‌عنوان مدل نمایش داده می‌شود.

روش SVM

SVM یک روش آماری غیرپارامتریک نظارت شده است که دوکلاس را با استفاده از یک مرز خطی از هم جدا می‌کند و سعی در پیدا کردن بهترین آبرصفحه‌ای دارد که بتواند با حداکثر فاصله، داده‌های مربوط به دو طبقه را از هم تفکیک کند. در این روش ابتدا نمونه‌هایی که مرز کلاس‌ها را تشکیل می‌دهند به دست می‌آیند. به عبارت دیگر تعدادی از نقاط آموزشی که کم‌ترین فاصله را تا مرز تصمیم‌گیری دارند می‌توانند به‌عنوان بردار پشتیبان در نظر گرفته شوند. در این روش اگر در فضای طیفی کلاس‌ها تداخل داشته باشند، داده‌ها به فضایی با ابعاد بالاتر نگاشت می‌شوند تا امکان تمایز آن‌ها میسر گردد. به منظور بردن داده‌ها به فضای با ابعاد بالاتر از توابعی موسوم به کرنل استفاده می‌شود. یک روش انتخاب کرنل مناسب، استفاده از روش‌های مجموعه ارزیاب می‌باشد. در این روش مدل SVM را با استفاده از کرنل‌ها یا پارامترهای مختلف، آموزش می‌دهند و هر کدام از پارامترها یا کرنل‌ها که دقت بهتری روی مجموعه داده ارزیاب داشته باشد به عنوان پارامتر و کرنل بهینه انتخاب می‌شوند. از آن-جایی که توابع کرنل زیادی وجود دارند که هر یک دارای خصوصیات مربوط به خود هستند و هیچ‌گونه شرط پیش-فرضی برای انتخاب تابع کرنل بهتر وجود ندارد (Zöller et al. 2013). در این پژوهش چند کرنل متفاوت روی داده‌ها اجرا شده و عملکرد هر کدام مورد ارزیابی قرار گرفت. کرنل‌هایی که در این پژوهش مورد بررسی قرار گرفتند عبارتند از: ضرب داخلی، چند جمله‌ای و شعاعی.

در روش SVM بسته به نوع کرنلی که استفاده می‌شود، تعیین مقدار مناسب برای برخی از پارامترها به افزایش دقت این روش در تشخیص کمک می‌کند.

یک مزیت SVM این است که برای کاربردهایی که تعداد برچسب کلاس در آن توازن و تعادل ندارد، مناسب است و چون در عملیات تشخیص پرت هم دقیقاً تعداد برچسب کلاس‌ها متعادل نیستند می‌تواند خوب عمل کند.

روش رگرسیون

خط رگرسیون ابزاری برای پیش‌بینی مقدار یک متغیر وابسته بر حسب متغیر مستقل است. در واقع برای مدل‌سازی مقادیر دو صفت خاصه، خطی را می‌یابد که به هم‌ه‌ی زوج مقادیر این دو صفت خاصه نزدیک باشد. روش‌های رگرسیون زمانی

در این روش هر یک از مدل‌ها با وزن مخصوصی در رأی-گیری شرکت می‌کنند (Zöller et al. 2013) و برای رأی-گیری نهایی مجموع وزن‌های هر مدلی که کلاسی را تخمین می‌زند، محاسبه می‌شود.

ساختار شبکه حسگر بی‌سیم پیشنهادی

خوشه‌بندی یک راه مؤثر برای بهینه‌سازی مصرف انرژی و در نتیجه بالا بردن طول عمر شبکه حسگر می‌باشد. با خوشه‌بندی، گره‌های شبکه حسگر به خوشه‌های مختلف تقسیم شده که هر خوشه توسط یکی از گره‌ها به نام سرخوشه مدیریت می‌شود و گره‌های عضو خوشه داده‌های خود را به سرخوشه می‌فرستند. در این ارسال اطلاعات ممکن است انرژی سرخوشه‌ها تحلیل رود، بنابراین سرخوشه‌هایی که انرژی باقیمانده‌ی کم‌تری دارند، به گره‌ی معمولی تبدیل می‌شوند و سپس یک گره با انرژی باقی‌مانده بالاتر به عنوان سرخوشه انتخاب خواهد شد. هدف اصلی الگوریتم‌های خوشه‌بندی حداقل کردن مصرف انرژی و افزایش طول عمر شبکه است. در این پژوهش به مسئله‌ی خوشه‌بندی در شبکه حسگر نیز پرداخته شده به گونه‌ای که از طریق مدیریت جریان داده سعی خواهد شد، پس از این که خوشه‌های شبکه شکل‌گرفته و ارسال داده از گره‌های زیرخوشه به سرخوشه صورت گرفت پس از مدتی شبکه همگرا شده و داده‌های ارسالی تا حدودی به صورت متناوب و با توزیع یکسانی صورت گیرد، بنابراین گره سرخوشه داده‌های هر گام را بررسی کرده و سعی در پالایش داده‌ها می‌کند و در صورتی که داده را پرت تشخیص دهد ارسال داده را به ایستگاه پایه گزارش خواهد کرد در غیر این صورت ارسال به صورت پیرودی و با نرخ ثابت ۱۰ دقیقه خواهد بود.

هم‌چنین روش ارسال در شبکه به صورت ترکیبی از ارسال تک‌گامی و چندگامی است. از آنجایی که در ارسال تک‌گامی انرژی زیادی از هر گره صرف انتقال داده به صورت مستقیم به گره مرکزی می‌شود، در نتیجه انرژی باتری در حسگرها خیلی زود تخلیه شده و طول عمر شبکه کاهش می‌یابد.

از سوی دیگر در ارسال چندگامی هر گره حسگر با کمک سایر حسگرهایی که در مسیر قرار دارند داده‌ها را به گره مرکزی می‌فرستند که باعث می‌شود بیشتر انرژی آن‌ها برای ایجاد ارتباط بین حسگرها مصرف شود. بنابراین بهترین

مناسب هستند که مقادیر صفت‌خاصه در مجموعه داده‌ها به برچسب کلاس بستگی داشته باشند. یکی از مدل‌های خطی تعمیم‌یافته رگرسیون لجستیک است که یک مدل آماری رگرسیون برای متغیرهای وابسته دو مقداره می‌باشد. منظور از دو مقداره بودن، رخ داد یک واقعه‌ی تصادفی در دو موقعیت ممکن است. در این پژوهش به دلیل دو مقداره بودن برچسب کلاس از روش رگرسیون لجستیک برای پیشگویی احتمال پرت بودن داده‌ها استفاده شده و از لحاظ معیارهایی مانند صحت، حساسیت، میانگین هارمونیک و هم‌چنین سرعت اجرا مورد ارزیابی و مقایسه قرار گرفت.

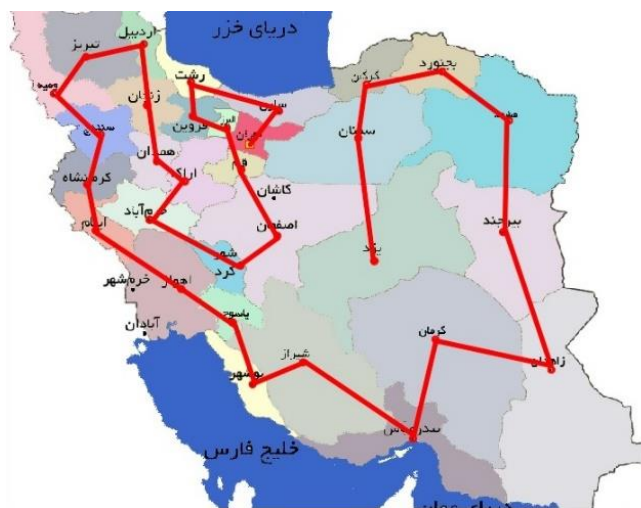
روش ترکیبی

یکی از راهکارهای موجود جهت افزایش دقت مدل استفاده ترکیبی از چند مدل به جای یک مدل است. با هدف بالا بردن دقت شناسایی داده‌های پرت در این پژوهش ترکیبی از روش‌های SVM و رگرسیون برای طبقه‌بندی داده‌ها به-کار گرفته شد. دو تکنیک شناخته‌شده برای ترکیب مدل‌ها bagging و boosting می‌باشد. در تکنیک bagging انتخاب داده‌ها برای آموزش مدل به صورت bootstrap انجام می‌شود که یک روش نمونه‌گیری با جایگزینی است و p مرتبه نمونه‌هایی را جهت آموزش یک مدل انتخاب کرده و با هر یک از این مجموعه داده‌های آموزشی و به کمک یک الگوریتم طبقه‌بندی مدلی ایجاد می‌شود و در نهایت p مدل مجزا وجود خواهد داشت. برای طبقه‌بندی یک نمونه جدید هر یک از مدل‌ها برچسب کلاس پیشنهادی را اعلام می‌کنند و رأی نهایی از طریق حداکثر تکرار برچسب کلاس می‌باشد (برچسب کلاسی که بیش‌تر پیشنهاد شده انتخاب می‌شود).

در تکنیک boosting برای انتخاب داده‌ها جهت آموزش مدل روش adaboost به کار گرفته می‌شود که داده‌های آموزشی توسط این تکنیک به p قسمت کاملاً مجزا تقسیم شده و توسط هر یک از این قسمت‌ها و با کمک الگوریتم طبقه‌بندی مورد نظر مدل ساخته می‌شود و با استفاده از داده‌های آزمایشی تست می‌شود و با توجه به نرخ خطا در مرحله بعد نمونه‌هایی که در پیش‌بینی برچسب کلاس آن‌ها احتمال خطای بیش‌تری وجود دارد با وزن بیش‌تری نسبت به سایر نمونه‌ها توسط مدل‌های دیگر بررسی می‌شوند. بنابراین

درحقیقت سرخوشه‌ی شهرستان مربوطه است، ارسال خواهند شد. گره‌ی مرکزی محلی پس از جمع‌آوری داده‌ها از خوشه‌های مختلف آن‌ها را به ایستگاه پایه‌ی محلی که سرخوشه‌ی استان است، ارسال کرده و درنهایت ایستگاه پایه‌ی محلی پس از جمع‌آوری داده‌ها از تمامی خوشه‌های خود که شهرستان‌های تابعه هستند، آن‌ها را به ایستگاه پایه-ی اصلی که درتهران قرار دارد ارسال می‌کند. خوشه‌بندی براساس الگوریتم پیشنهادی درسطح شهرستان انجام شده و خوشه‌هایی شکل می‌گیرند. برای هر یک از این خوشه‌ها یک سرخوشه طبق آنچه که در بخش بعد ذکر خواهد شد، انتخاب می‌شود که تمام اعضا دریک خوشه، ابتدا داده‌های حس شده را به این سرخوشه ارسال کرده و سرخوشه پس از تجمیع داده‌ها و حذف داده‌های تکراری و هم‌چنین پس از شناسایی داده‌های پرت، آن‌ها را به گره‌ی مرکزی محلی که سرخوشه‌ی محلی خوشه‌ی شهرستان است، ارسال می‌کند. گره‌ی مرکزی محلی هم وظیفه‌ی ارسال تمام داده‌های جمع‌آوری شده از خوشه‌های خود به ایستگاه پایه‌ی محلی در سطح استان را خواهد داشت. در شکل ۲ نمای کلی از این ستون فقرات مجازی نشان داده شده است.

روش خوشه‌بندی گره‌ها استفاده ترکیبی از ارسال تک‌گامی و چندگامی برای دستیابی به هدف کاهش مصرف انرژی در شبکه حسگرگی سیم است. شبکه‌ی حسگر مدنظر در این پژوهش، یک شبکه‌ی حسگرگی سیم سلسله‌مراتبی براساس خوشه‌بندی است، به-این‌صورت که در بالاترین سطح به تعداد هراستان یک خوشه وجود دارد و ایستگاه پایه‌ی اصلی درتهران قرار دارد. دریک سطح پایین‌تر، داخل خوشه‌ی هراستان، نیز به تعداد شهرستان‌های آن استان خوشه وجود خواهد داشت و در آخرین سطح هر یک از شهرستان‌ها نیز خود به‌صورت خوشه‌بندی شده هستند. به‌منظور خوشه‌بندی در آخرین سطح که مربوط به خوشه‌بندی داخل هر یک از شهرستان-هاست، گره‌های حسگر به‌صورت تصادفی در مناطق مختلف پخش شده و از طریق الگوریتم جدول ۱، عملیات خوشه-بندی حسگرها و انتخاب سرخوشه انجام می‌شود. هر یک از خوشه‌ها یک سرخوشه‌ی محلی دارند که عملیات جمع-آوری داده‌ها از سایر حسگرها، تجمیع و تشخیص داده‌ی پرت و اضافی در آن‌ها انجام شده و سپس به‌منظور انجام پردازش‌های بیش‌تر داده‌ها به گره‌ی مرکزی محلی که



شکل ۲ - ستون فقرات مجازی در سطح کشور

تشخیص تکراری در هر یک از گره‌ها و تشخیص پرت در سرخوشه انجام می‌شود.

خوشه‌بندی در پایین‌ترین سطح سلسله‌مراتب شبکه‌ی مذکور براساس الگوریتم زیر انجام می‌شود. عملیات مربوط به

جدول ۱- شبه کد الگوریتم پیشنهادی

Proposed Algorithm in clustering and Re clustering Phase:

```

For each node (i) do
{
Num-neighbors = find-neighbors (i)
If (Num-neighbors > 0){
P = create-vector (Num-neighbors)
Initialize-probabilities-vector (pi); // Select a node as a header
probability vector
j = select max-prob (pi)
Set-CH (j);
If (Num-neighbors (j) > Avg (Num-neighbors (neighbors(j))) )
Increase (pj)
If ( j is neighbor of CH-node)
Decrease (pj)
If(distance to BS(j)<distance to BS (neighbors(j)))
Increase (pj)
Else Decrease (pj)
If(j is cluster head & num-neighbors (cluster node(j)) / num-neighbors(CH(j))>N/ k )
Decrease (pj)

}
Else set-CH ()
}

```

Proposed Algorithm: finding Outlier Detection for maximum of life time for unicast routing

Input : A Stochastic Graph $G = (V, E, W)$

Assumptions:

Let V_i be selected node i th;

Let W_i be the remaining energy of V_i ;

Let D_i be the distance of sink of V_i ;

Let $E(n)$ be the vector for saving the remaining _ energy of selected node and initially is $\{0\}$

;

Let Vector be the vector for saving one hop neighbors of V_i ;

Let P_{end} is the threshold for ending Probability;

Let C be counter of cluster head members and initially is $\{0\}$;

Begin Algorithm

Repeat

Step1- <construct phase>

Let SN be Sink Node;

Let V_s be the source node of unicast routing ;

Start from V_s to create for unicast routing ;

list_determined= V_s ;

Nee_list = Neighbors (V_s) ;

$E[1]= W_s$;

$C=1$;

while

Select one of the neighbors from Nee_list with maximum neighbor of degree; // as V_i

list_labeled = list_labeled $\cup V_i$;

Nee_list = Nee_list \cup Neighbors (V_i) ;

$C= C+1$;

$E[c] = W_i$;

end while

// find one hop neighbors with max neighbor of degree and placed in list_labeled

step2- <Probability phase>

while // to determine list_determined

```

Select one if the elements list_labeled;
If ( $E_i > E_{avg}$ ) and ( $D_i < D_{avg}$ ) then
Increase probability;
else decrease probability;
If (number of neighbors of  $V_i >$  average number of its neighbor nodes) then
Increase probability;
else decrease probability;
list_determined = list_labeled  $\cup$   $V_i$  ;//based on max probability
Nee_list = Nee_list  $\cup$  Neighbors ( $V_i$  );
C= C+1 ;
E[c] =  $W_i$  ;
end while
if ( $V_i$  is Nee and hasn't neighbor of Head ) then
 $V_i$  will be one of elements list_determined;
until (Probability of selection  $> P_{end}$  ) ;
send the unicast message through the created route to sink node;
End Algorithm.

```

Procedure outlier detection

The global normal pattern (-1)

```

begin
for each time window do
for all member nodes  $\in$  do
Sense a set of data vectors () and normalize it
Apply the clustering procedure on () and send
the data () to the group head
end for
for group head do
Detect outliers based upon (-1)// outlier is data that  $> \mu \pm 3\sigma$ 
end for
end for
end procedure

```

Procedure data redundancy detection

```

begin
for each time window do //the window is allocated to each node follower
for all member nodes  $\in$  do
Sense a set of data vectors () and Check with previous data
Apply the Measure the distance
If Measure the distance  $>$  Defined threshold
send the data () to the cluster head
end for
the next time window //next node follower
end for
end procedure

```

ابتدا با احتمال اولیه یکسان، گره‌های حسگر با یک احتمال بر مبنای انرژی گره‌ها یکی از حالت‌های سرخوشه یا معمولی را انتخاب می‌کنند و در نهایت هر گرهی معمولی سرخوشه‌ی خود را شناسایی می‌نماید و خوشه‌ها تشکیل می‌شوند. سپس فاز ارسال داده شروع می‌گردد که در فاز ارسال داده، گره‌های معمولی بر اساس برنامه زمان‌بندی

الگوریتم پیشنهادی شامل چهار فاز می‌باشد: فاز شناسایی، فاز خوشه‌بندی اولیه، فاز ارسال داده و فاز خوشه‌بندی مجدد؛ در فاز شناسایی بعد از پراکنده شدن گره‌های حسگر، هریک از آن‌ها اقدام به شناسایی گره‌های همسایه خود می‌نمایند. در این فاز بر مبنای انرژی گره‌های اولیه و فاصله، جدول همسایگی تشکیل می‌دهند. در فاز خوشه‌بندی اولیه

می‌کند. تا مادامی که انرژی گرهی سرخوشه انتخابی از میانگین انرژی باقی‌مانده گره‌های عضو خوشه بیش‌تر باشد وارد فاز خوشه بندی مجدد نخواهد شد.

با توجه به این‌که روش پیشنهادی شناسایی و حذف داده‌ی پرت بوده، گزارشی که به ایستگاه پایه ارسال خواهد شد مبنی بر پرت بودن داده حسگرهای زیرمجموعه این خوشه بوده که چنانچه ایستگاه پایه تشخیص دهد آن را یا حذف خواهد کرد و یا برای آنالیز بیشتر مورد داده‌کاوی زمانی قرار خواهد داد. این رویکرد باعث می‌شود علاوه بر این‌که شبکه‌ی حسگر به گونه‌ای خوشه‌بندی شود که انرژی به صورت یکنواخت در شبکه به‌مصرف رسیده و طول عمر شبکه افزایش یابد، مدیریت ارسال داده نیز به‌گونه‌ای مطلوب صورت گیرد و از ارسال غیر ضروری داده در شبکه جلوگیری شود و انرژی نسبت به روش‌های مشابه ذخیره گردد.

ارزیابی

این پژوهش روی داده‌های یک‌سال ایستگاه خودکار اداری هواشناسی انجام شده که توسط حسگرهای این مجموعه جمع‌آوری شده‌اند. حدود ۲۷۰۰۰ نمونه داده که دارای ۶ صفت خاصه اعم از تاریخ و زمان، دمای هوا، فشار هوا، میزان رطوبت، سمت و سرعت باد بوده و برچسب کلاس این داده‌ها پرت یا نرمال می‌باشد. داده‌های پرت با برچسب ۱ و داده‌های نرمال با برچسب ۰ مشخص شده‌اند.

همان‌طور که در شکل‌های ۳ و ۴ مشاهده می‌شود روش‌های SVM، رگرسیون و ترکیبی به‌ازای مقادیر متفاوت k اجرا شد. سپس مقادیری از k که به‌ازاء آن، روش اجرا شده روی داده‌ها نتایج بهتری از لحاظ معیارهای ارزیابی داشته به‌عنوان مقدار مناسب k در نظر گرفته شد. کرنل یک معیار شباهت بین داده‌هاست انتخاب کرنل‌های غیرخطی اجازه ی ساخت جداکننده‌های خطی در فضای ویژگی را می‌دهد در صورتی که آنها در فضای اصلی غیرخطی هستند.

در هر نوع داده‌ای مثل داده‌های هواشناسی می‌توان بر اساس ویژگی‌های دامنه مورد نظر یک کرنل مناسب تعریف کرد و از ماتریس کرنل داده‌های آموزش که با هر روش

که سرخوشه برای آن‌ها تعیین نموده است، داده‌های جمع-آوری شده را به صورت تناوبی به سرخوشه ارسال می‌دارند و سرخوشه نیز اطلاعات را به گرهی مرکزی می‌فرستد. در این فاز به منظور جلوگیری از ارسال داده‌های تکراری گره‌های پیرو داده‌های جدید را براساس معیار فاصله‌ی مینکوفسکی با داده‌های قبلی مورد بررسی قرار می‌دهند. اگر اختلاف داده‌های جدید با داده‌های قبلی از یک آستانه‌ای که به صورت دستی برای سیستم تعیین شده بیش‌تر شد، داده را ارسال می‌کند، در غیراین‌صورت آن را به‌عنوان داده تکراری شناسایی کرده و از ارسال داده به گرهی سرخوشه جلوگیری می‌کند و نوبت خود را به گره بعدی برای ارسال به سرخوشه می‌دهد. اضافه کردن این قسمت به فاز ارسال داده از این جهت دارای اهمیت است که گره سرخوشه به علت این‌که خود نیز نیاز به ارسال تک‌گامی بر روی ستون فقرات مجازی دارد بیش‌تر در معرض کاهش سریع انرژی و در نتیجه افراز شبکه است پس بهتر است بارارسالی داده تکراری بر روی سرخوشه را با حذف کردن داده‌های تکراری کاهش دهیم. در این فاز که بخش نوآوری روش پیشنهادی می‌باشد داده‌هایی که نویز و یا داده پرت می‌باشند، شناسایی می‌شوند. روش شناسایی با توجه به نوع داده‌های هواشناسی که تقریباً توزیع متوازن و مشخصی دارند به این صورت است که داده‌هایی که اغلب سه یا بیش از سه واحد انحراف معیار از میانگین توزیع مربوط به خودشان را دارند به عنوان داده پرت مشخص شده است و براساس این داده‌ها و با کمک تکنیک‌های طبقه‌بندی مذکور در ابتدا مدلی برای تشخیص داده‌ی پرت در سرخوشه ایجاد شده و با استفاده از آن داده‌های پرت شناسایی می‌شوند؛ و بدین شکل با مدیریت داده هم از ارسال بیش از حد داده به گره مرکزی جلوگیری شده و هم به ثبت و نگهداری داده‌های پرت پرداخته شده چرا که ممکن است تحلیل خود این داده‌ها ارزشمند باشد.

گره‌های سرخوشه به دلیل مصرف زیاد انرژی زود تخلیه شده و از بین می‌روند به همین دلیل در فاز خوشه‌بندی مجدد، هرگاه میزان انرژی گرهی سرخوشه از حد مشخصی پایین‌تر بیاید، گرهی سرخوشه یک گرهی دیگر از همان خوشه که انرژی بیش‌تری دارد را به‌عنوان سرخوشه انتخاب

طورکه در شکل ۵ نشان داده شده، تکنیک SVM با چند کرنل روی داده‌ها اجرا شده و براساس ارزیابی‌های صورت‌گرفته کرنل چندجمله‌ای از درجه ۳ نسبت به سایر توابع کرنل روی این داده‌ها عملکرد بهتری را نشان می‌دهد. در نهایت طبق آنچه که در شکل ۷ مشاهده می‌شود؛ روش‌های SVM، رگرسیون لجستیک و تکنیک‌های ترکیبی Adaboost و Bootstrap با مقادیر بهینه برای پارامترها روی داده‌ها اجرا شدند.

ممکن به دست آمده است یک مدل ماشینی بردار پشتیبان را آموزش داد. همچنین تکنیک SVM در این پژوهش با ۳ کرنل چندجمله‌ای، ضرب داخلی و شعاعی روی داده‌ها اجرا شد. براساس معیارهای ارزیابی به دست آمده با توجه به شکل ۵، کرنل چندجمله‌ای از درجه ۳ و با توجه به شکل ۶، کرنل شعاعی با ۰.۳ برابر ۷ به عنوان مقادیر بهینه برای این پارامترها تشخیص داده شدند. با هدف انتخاب کرنل مناسب همان-

مقایسه مقادیر K برای تکنیک SVM					مقایسه مقادیر K برای تکنیک رگرسیون				
F-score	صحت	حساسیت	دقت	K مقادیر	F-score	صحت	حساسیت	دقت	K مقادیر
0.84	99.64%	82.00%	85.42%	2	0.91	99.80%	86.00%	95.56%	2
0.90	99.78%	90.00%	90.00%	3	0.97	99.93%	96.00%	97.96%	3
0.93	99.84%	96.00%	90.57%	5	0.97	99.93%	98.00%	96.08%	5
0.90	99.78%	90.00%	90.00%	7	0.97	99.93%	96.00%	97.96%	7
0.92	99.82%	92.00%	92.00%	8	0.97	99.93%	96.00%	97.96%	8
0.92	99.82%	94.00%	90.38%	10	0.95	99.89%	92.00%	97.87%	10

شکل ۳ - مقایسه مقادیر متفاوت برای پارامتر k در روش‌های SVM و رگرسیون لجستیک به منظور یافتن مقادیر بهینه

مقایسه مقادیر K برای تکنیک ترکیبی				
F-score	صحت	حساسیت	دقت	K مقادیر
0.91	99.80%	86.00%	95.56%	2
0.97	99.93%	96.00%	97.96%	3
0.97	99.93%	98.00%	96.08%	5
0.97	99.93%	96.00%	97.96%	7
0.97	99.93%	96.00%	97.96%	8
0.97	99.93%	96.00%	97.96%	10

شکل ۴ - مقایسه مقادیر متفاوت برای پارامتر k در روش ترکیبی به منظور یافتن مقادیر بهینه

مقایسه مقادیر مختلف برای پارامتر گاما در کرنل شعاعی					مقایسه درجات مختلف کرنل چندجمله‌ای				
F-score	صحت	حساسیت	دقت	مقادیر گاما	F-score	صحت	حساسیت	دقت	درجه
0.83	99.62%	84.00%	82.35%	0.3	0.91	99.80%	94.00%	88.68%	2
0.80	99.55%	82.00%	78.85%	0.5	0.92	99.82%	94.00%	90.38%	3
0.76	99.44%	78.00%	73.58%	0.8	0.91	99.80%	92.00%	90.20%	4
0.32	96.35%	78.00%	20.42%	1	0.90	99.78%	90.00%	90.00%	5
0.78	99.53%	74.00%	82.22%	2					

شکل ۵ - مقایسه درجات مختلف کرنل چندجمله‌ای، شکل ۶ مقایسه مقادیر مختلف پارامتر گاما در کرنل شعاعی

مقایسه کرنل‌ها برای روش SVM						
نوع کرنل	دقت	حساسیت	صحت	F-score	حاشیه	نرخ خطا
چندجمله‌ای	90.38%	94.00%	99.82%	0.922	0.478	0.18%
ضرب داخلی	92.00%	92.00%	99.82%	0.920	0.451	0.18%
شعاعی	79.63%	86.00%	99.60%	0.827	0.421	0.40%

شکل ۷ - مقایسه اجرای SVM با کرنل‌های مختلف

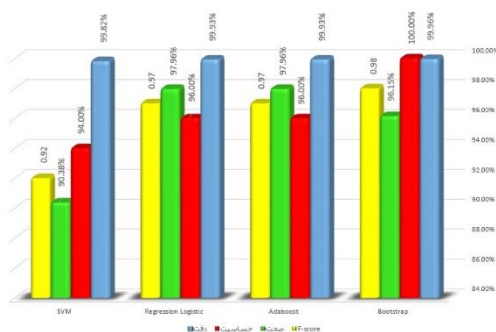
روی نتایج آن رگرسیون اجرا گردد، عملکرد بهتری از لحاظ معیارهای اندازه‌گیری شده خواهد داشت. تکنیک‌های مورد استفاده در این پژوهش براساس مقادیر بهینه پارامترها روی داده‌ها اجرا گردیده و نتایج آن با هدف یافتن

از آنجایی که ترتیب اجرای تکنیک‌ها در روش ترکیبی روی نتیجه حاصل شده در تشخیص پرت تأثیر دارد، در تکنیک‌های ترکیبی، ترتیب اجرای الگوریتم‌ها در هر یک از روش‌ها مورد بررسی قرار گرفت. همان‌طور که در شکل ۸ نشان داده می‌شود درحالتی که ابتدا SVM روی داده‌ها اجرا شده و سپس

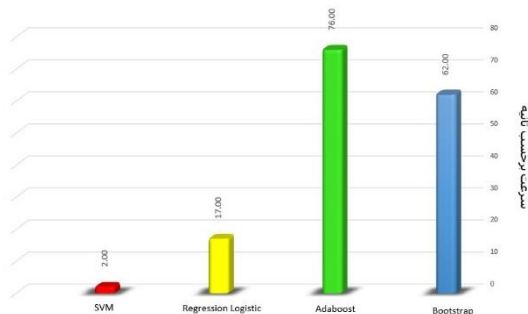
روشی مناسب از لحاظ سرعت، دقت و ... جهت تشخیص داده پرت، بایکدیگر مقایسه شدند. باتوجه به شکل‌های ۹ و ۱۰ از بین روش‌های پیشنهادی تکنیک SVM با کرنل چندجمله‌ای از درجه ۳ سرعت بالاتری را در کنار عملکرد مطلوب به همراه دارد.

مقایسه تکنیک‌های ترکیبی از لحاظ ترتیب اجرای الگوریتم‌ها				
F-score	صحت	حساسیت	دقت	تکنیک‌های ترکیبی
0.91	99.80%	90.00%	91.84%	Bootstrap(Regression+Svm)
0.98	99.96%	100.00%	96.15%	Bootstrap(Svm+Regression)
0.92	99.82%	92.00%	92.00%	Adaboost(Regression+Svm)
0.97	99.93%	96.00%	97.96%	Adaboost(Svm+Regression)

شکل ۸ - مقایسه ترتیب اجرای الگوریتم‌ها در تکنیک‌های ترکیبی



شکل ۹ - مقایسه تکنیک‌های پیشنهادی با مقادیر بهینه پارامترها



شکل ۱۰ - مقایسه تکنیک‌های پیشنهادی از لحاظ سرعت اجرا

توسط آنها، سراسری و یکتا می باشد قاعدتا این روش با کرنل‌های مختلف ارزیابی شد.

نتیجه‌گیری و کارهای آینده

هدف در این پژوهش، انتخاب مقادیر بهینه برای پارامترها و ساختن مدلی که بتواند بالاترین دقت را در شناسایی داده‌ی پرت در کم‌ترین زمان ممکن داشته باشد، به منظور کاهش مصرف انرژی در شبکه‌ی حسگر بی‌سیم می‌باشد. برای این منظور از تکنیک‌های طبقه‌بندی SVM، رگرسیون و ترکیبی Adaboost و Bootstrap برای شناسایی داده‌های پرت استفاده شد و با استفاده از این تکنیک‌ها در شبکه‌های

به عبارت ساده روش‌های مختلف SVM محدودیت‌های ذاتی دارند به عنوان مثال هنوز مشخص نیست که به ازای یک تابع نگاشت، پارامترها را چگونه باید تعیین کرد و معمولاً با آزمون و خطا مقادیر مختلف در کاربردهای مشخص می‌توان بهترین پارامتر را پیدا کرد. همچنین SVM محاسبات پیچیده و زمان بر نیاز دارند و به دلیل پیچیدگی محاسباتی، حافظه زیادی نیز مصرف می‌کنند ولی با توجه به ذات مسأله‌ی مورد هدف در این پایان‌نامه چون SVM‌ها دارای یک شالوده نظری منسجم بوده و جواب‌های تولید شده

10. Esmaeili ,M .(2013). Training of step by step data mining with Rapied Miner. Tehran : Atynegar publishing, pp. 134. (In Farsi)
11. Fawzy, A., Mokhtar, H.M.O., Hegazy, O.(2013). Outlier detection and classification in wireless sensor networks. Egyptian Informatics Journal- Elsevier, Vol 14, July 2013, pp. 157–164.
12. Gil, P., Martins, H., Januário, F.(2016). Detection and accommodation of outliers in wireless sensor networks within a multi-agent framework. Applied Soft Computing- Elsevier, Vol 42, May 2016, pp.204–214.
13. Gupta, G., Younis, M.(2003). Fault-tolerant clustering of wireless sensor networks. Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE, 16-20 March. 2003.
14. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.(2000). Energy-efficient communication protocol for wireless microsensor networks. System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, 7-7 Jan. 2000.
15. Hansen, E., Neander, J., Nolin, M., Bjorkman, M.(2006). Efficient cluster formation for sensor networks. The Online Publication ..., 2006 .
16. Janakiram ,D., Mallikarjuna, A., Reddy, V., Kumar, P.(2006). Outlier detection in wireless sensor networks using bayesian belief networks. Communication System Software and Middleware, 2006. Comsware 2006. First International Conference on, 8-12 Jan. 2006.
17. Kiss, F.(2003). Credit scoring processes from a knowledge management perspective. Periodica Polytechnica. Social and Management Sciences 11.1 , 2003, pp. 95-110.
18. Li, M .(2014). Data filtering and distortion-prevention for power saving in wireless sensor networks. Communication Problem-Solving (ICCP), 2014 IEEE International Conference on Beijing, 5-7 Dec. 2014 , pp. 103-106.
19. Mhatre, V., Rosenberg, C.(2004). Design guidelines for wireless sensor networks: communication, clustering and aggregation. Ad Hoc Networks- Elsevier, Vol 2, January 2004, pp. 45–63.
20. Pantazis, N.A., Nikolidakis, S.A., Vergados, D.D.(2013). Energy-efficient routing protocols in wireless sensor networks: a survey. IEEE Communications Surveys & Tutorials, Vol 15, No 1, March 2013, pp. 103–127.
- حسگر بی‌سیم که از پروتکل سلسله‌مراتبی برای مسیریابی استفاده می‌کنند به همراه ستون فقرات مجازی ، سعی شد مصرف انرژی در این شبکه‌ها کاهش یابد. نتایج آزمایشات نشان داد روش پیشنهادی با فیلتر کردن داده های پرت و تکراری از ارسال بیهوده داده جلوگیری کرده و همچنین طول عمر شبکه را افزایش داده است.
- ### منابع
1. Asgarnezhad, R., Akbari Torkestani, j.(2011). A survey on backbone formation algorithms for wireless sensor networks(a new classification). Telecommunication Networks and Applications Conference (ATNAC), 2011 Australasian, 9-11 Nov. 2011. (In Farsi)
 2. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cyirci, E .(2002). Wireless sensor networks: a survey. Computer Networks, 2002- Elsevier, Vol 38, 15 March. 2002, pp. 393–422.
 3. Akkaya, K., Younis, M.(2005). A survey on routing protocols for wireless sensor networks. Ad Hoc Networks- Elsevier, Vol 3, May 2005, pp. 325–349.
 4. Bastam , M . (2008). Data collection on wireless sensor networks in order to maximize your lifetime. Master's thesis, Department of Computer Engineering and Information Technology, Amirkabir University of Technology. (In Farsi)
 5. Bandyopadhyay, S., Coyle, E.J.(2003). An energy efficient hierarchical clustering algorithm for wireless sensor networks. INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies, 30 March-3 April. 2003.
 6. Bamnett, V., Lewis, T.(1994). Outliers in statistical data. Journal of the Royal Statistical Society. Series A (General), Vol. 141, No. 4 , pp. 1-2.
 7. Corke , P., Wark, T., Jurdak, R., Hu, W., Valencia, P., Moore, D.(2010). Environmental wireless sensor networks. Proceedings of the IEEE, Vol 98, Nov. 2010 , pp. 1903 – 1917.
 8. Devika, R., Santhi, B., Sivasubramanian, T. (2013). Survey on routing protocol in wireless sensor network. International Journal of Engineering and Technology, Vol 5, No 1, Feb-Mar 2013. Pp.350-356.
 9. Esmaeili ,M .(2016). Data warehouses and data mining. Tehran: Cultural and Art Institute of Dybagaran , pp. 416. (In Farsi)

- Surveys & Tutorials , Vol13, Third Quarter 2011, pp. 443 – 461.
28. Tsai, C.W., Lai, C.F., Vasilakos, A.V.(2014). Future internet of things: open issues and challenges. *Wireless Networks-Springer*, Vol 20, November 2014, pp. 2201–2217.
 29. Vy ,V.D., An, T.H., Tu ,B.D., Nghe, V.D., Du, H.Q.(2012). Data reduction algorithms for wireless sensor networks in environment monitoring and warning applications. *Computational Intelligence, Modelling and Simulation (CIMSIM)*, 2012 Fourth International Conference on Kuantan, 25-27 Sept , pp. 416-421.
 30. Wang, L.C., Wang, C.W.(2004). A cross-layer design of clustering architecture for wireless sensor networks. *Networking, Sensing and Control*, 2004 IEEE International Conference on, 21-23 March. 2004.
 31. Xu, S. , Hu, C. , Wang, L. , Zhang, G .(2012). Support vector machines based on k nearest neighbor algorithm for outlier detection in wsns . *Wireless Communications, Networking and Mobile Computing (WiCOM)*, 2012 8th International Conference on Shanghai , 21-23 Sept , pp.1-4.
 32. Younis, O., Fahmy, S.(2004). Distributed clustering in ad-hoc sensor networks: a hybrid, energy-efficient approach. *INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies*, 7-11 March .2004.
 33. Zöller, S., Vollmer, C., Wachtel ,M., Steinmetz, R. , Reinhardt, A.(2013). Data filtering for wireless sensor networks using forecasting and value of information. *Local Computer Networks (LCN)*, 2013 IEEE 38th Conference on, 21-24 Oct. 2013.
 34. Zhang Y, Meratnia N, Havinga P.(2010). Outlier detection techniques for wireless sensor networks: a survey. *IEEE Communications Surveys & Tutorials* , Vol12, Second Quarter 2010 ,pp. 159 – 170
 - Tutorials, Vol15, Second Quarter 2013 ,pp. 551 – 591.
 21. Raghuwanshi, S., Mishra, A.(2003). A self-adaptive clustering based algorithm for increased energy-efficiency and scalability in wireless sensor networks. *Vehicular Technology Conference*, 2003. VTC 2003-Fall. 2003 IEEE 58th, 6-9 Oct. 2003.
 22. Rezaei, Z., Akbari Torkestani, J.(2012). An Energy-efficient mcdfs-based routing algorithm for wireless sensor networks: learning automata approach. *Przegląd Elektrotechniczny (Electrical Review)*, 2012. (In Farsi)
 23. Shukla, D.S, Pandey, A.C , Kulhari, A. (2014) . Outlier detection: a survey on techniques of wsns involving event and error based outliers . *International Conference on Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH14)*. 28 & 29 November 2014, pp. 113 – 116.
 24. Shahid, N., Naqvi, I.H. , Qaisar, S.B.(2013). One-class support vector machines: analysis of outlier detection for wireless sensor networks in harsh environments. *Artificial Intelligence Review* , vol 43, no. 4: 515-563.
 25. Shen ,X., Li, Z., Jiang, Z., Zhan, Y. (2013). Distributed svm classification with redundant data removing . *2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing* , Beijing, 20-23 Aug , pp. 866-870.
 26. Sepasi zahmati , A. (2006). Design, simulation and evaluation of a new protocol to improve network routing of wireless sensor networks. Master's thesis, Department of Electrical Engineering, University of Science and Technology. (In Farsi)
 27. Sudevalayam, S., Kulkarni, P.(2011). Energy harvesting sensor nodes: survey and implications. *IEEE Communications*