
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گروه هوش مصنوعی

یک الگوریتم کارآ برای همجوشی داده ها در شرایط عدم قطعیت

محمدمهدی علیان نژادی

استاد راهنما

دکتر علی اکبر پویان

استاد مشاور

دکتر وحید ابوالقاسمی

پایان نامه جهت اخذ درجه کارشناسی ارشد

تابستان ۹۳

دانشکده : مهندسی کامپیوتر و فن‌آوری اطلاعات

گروه : هوش مصنوعی

پایان نامه کارشناسي ارشد آقاي محمدمهدی علیان‌نژادی

تحت عنوان: يك الگوريتم كارآ براي همجوشی داده ها در شرایط عدم قطعیت

در تاريخ ..... توسط كميته تخصصي زير جهت اخذ مدرک كارشناسي ارشد مورد ارزيابي و با درجه ..... مورد پذيرش قرار گرفت.

اساتید راهنما	امضاء	اساتید مشاور	امضاء
نام و نام خانوادگی :	**دکتر علی اکبر پویان**	نام و نام خانوادگی :	**دکتر وحید ابوالقاسمی**
نام و نام خانوادگی :		نام و نام خانوادگی :	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی :		نام و نام خانوادگی :	
نام و نام خانوادگی :			
نام و نام خانوادگی :			
نام و نام خانوادگی :			

تقدیم به...

خدایی که آفرید

جهان را، انسان را، عقل را، علم را، معرفت را،

عشق را

و به کسانی که عشقشان را در وجودم دمید.

شکر شایان نثار ایزد منان که توفیق را رفیق راهم ساخت تا این پایان نامه را به اتمام برسانم.

برخود لازم می‌دانم تا از همه‌ی کسانی که مرا در پیشبرد این پایان نامه یاری نمودند، سپاس و قدرانی به عمل آورم

ابتدا از زحمات اساتید فاضل و اندیشمند جناب آقای دکتر پویان به عنوان استاد راهنما و جناب آقای دکتر ابوالقاسمی به عنوان استاد مشاور که مرا صمیمانه و مشفقانه یاری داده‌اند، کمال تشکر را دارم.

تشکر و سپاس از اساتید دانشمند و پر مایه ام جناب آقای دکتر حسن‌پور، جناب آقای دکتر زاهدی و خانم دکتر مشایخی که از محضر پرفیض تدریسشان، بهره‌ها برده‌ام.

و نیز سپاسگذارم از خدمت پدر و مادر بسیار عزیز، دلسوز و فداکارم که پیوسته جرعه نوش جام تعلیم و تربیت فضیلت و انسانیت آنها بوده‌ام و همواره چراغ وجودشان روشنگر راه من در سختی‌ها و مشکلات بوده است.

علاوه بر این بر خود واجب میدانم می‌دانم از همسرم که مسیح وار با صبرش در تمامی لحظات رفیق راهم بود، سپاس و قدردانی به عمل آورم.

تعهد نامه

اینجانب محمدمهدی علیان‌نژادی دانشجوی دوره کارشناسی ارشد رشته هوش مصنوعی دانشکده مهندسی کامپیوتر و فن‌آوری اطلاعات دانشگاه صنعتی شاهرود نویسنده پایان نامه یک الگوریتم کارآ برای همجوشی داده ها در شرایط عدم قطعیت تحت راهنمائی دکتر علی اکبر پویان متعهد می شوم .

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

تاریخ **امضای**

دانشجو

چکیده

در این پایان نامه همجوشی داده ها در شرایط عدم قطعیت به صورت "ترکیب چند منبع غیر قطعی با توانایی شناسایی و برخورد با عدم قطعیت و تناقضات که بتواند منجر به ارائه یک بازنمایی مؤثر برای انسان یا ماشین شود" تعریف شده است. مهمترین جنبه این پایان نامه شناسایی و رسیدگی به عدم قطعیت و تناقضات داده ها در همجوشی برای سیستم‌های ناشناس می باشد. غالبا روش های پیشین همجوشی داده ها مانند فیلتر کالمن و تئوری بیزین با مدل رفتاری سیستم کار می کنند؛ این بدین معنا است که برای همجوشی داده ها از یک پیش بینی نظری استفاده می کنند. در سیستم هایی که مدل رفتاری آنها در دسترس نیست نمی‌توان از الگوریتمی مانند فیلتر کالمن استفاده نمود. باید به این موضوع توجه نمود که الگوریتم ارائه شده در این پایان نامه، می تواند برای سیستم با مدل مشخص کار نماید؛ اما بدیهی است با حذف مدل سیستم فرآیند همجوشی و رسیدگی به خطاها مشکل تر خواهد شد. مهمترین تفاوتی که در سیستم هایی با مدل رفتاری ناشناس و شناسا وجود دارد این است که در سیستم های شناسا، عملیات همجوشی در شرایط عدم قطعیت حتی با یک سنسور هم امکان پذیر است زیرا مدل سیستم در تولید داده همجوشی شده، کمک خواهد کرد. در سیستم های ناشناس برای رسیدگی به تناقضات به تعداد زیادی سنسور نیاز خواهد بود. اگر منابع دارای عدم قطعیت باشند، ممکن است همجوشی داده های آنها از واقعیت دور بوده و ارزیابی سیستم از محیط اشتباه باشد. روش ارائه شده در این پایان نامه، این امکان را کاهش می دهد. روش ارائه شده در این پایان نامه، شامل یک تکنیک جدید خوشه بندی، شبکه عصبی MLP و قانون جدید به روز رسانی پیش بینی برای همجوشی داده ها می باشد. در نهایت روش ارائه شده با داده های هواشناسی مورد آزمایش قرار گرفته و نتایج آزمایشات بیانگر توانایی روش ارائه شده می‌باشد.

کلمات کلیدی: همجوشی داده ها، عدم قطعیت، خوشه بندی، شبکه عصبی، پیش بینی وضع هوا

فهرست مطالب

فهرست شکل ها

فهرست جدول ها

فصل اول – مقدمه و کلیات

۱– ۱– تعریف همجوشی داده ها

همجوشی داده ها یک تکنیک پیشرفته برای ترکیب داده های چند منبع به منظور بدست آوردن جواب‌های دقیق‌تر می باشد. برای مثال مغز انسان به کمک حس های لامسه، بینایی، بویایی، چشایی و شنوایی یک استنتاج از محیط انجام می دهد. به عنوان مثال دیگر، می‌توان توانایی بینایی در انسان را مورد بررسی قرار داد. هنگامی که با دو چشم یک تصویر را می بینید همجوشی تصاویر دیده شده توسط هر چشم در ذهن شما تداعی می‌شود. برای آزمایش این موضوع کافی است یک شیء را در جلو چشمان خود قرار دهید. سپس یک چشم خود را ببندید و تنها با چشم دیگر شیء را ببینید؛ حال بدون جا به جا نمودن شیء چشم باز خود را بسته و دیگری را باز نمایید؛ حال هر دو چشم را باز نمایید. همانطور که مشاهده خواهید نمود سه تصویر متفاوت از شیء دیده می‌شود.

همجوشی داده ها ابزار مفیدی برای رباتیک، پردازش سیگنال، شبکه های سنسوری و ... می‌باشد. هسته و پایه‌ی همجوشی داده ها آنچنان معروف و پرکاربرد می باشد که این تکنیک در تمام زندگی ما داخل شده است؛ بنابراین هر یک از محققین در رشته های متفاوت، از این تکنیک با اصطلاحات خود استفاده می نمایند.

همجوشی داده ها یک پژوهش چند رشته ای می باشد که ایده هایی از رشته های مختلف مانند پردازش سیگنال، تئوری اطلاعات، تخمین آماری، استنتاج و هوش مصنوعی گرفته است. [1]

تعاریف زیادی برای همجوشی داده ها وجود دارد. کلین در کتاب خود، همجوشی داده ها را به عنوان تعیین وضعیت داده‌های که توسط یک یا چند منبع تولید شده اند، تعریف می‌نماید [2] . در مقاله [3] همجوشی داده‌ها به عنوان "تبدیل اتوماتیک یا نیمه اتوماتیک داده ها از چند منبع

و در زمان متفاوت به یک بازنمایی به منظور کمک به تصمیم گیری انسان یا تصمیم گیری خودکار" تعریف کرده است.

۱–۲– عدم قطعیت و تناقض در داده ها

به طور کلی، هر منبع داده ای می تواند شامل سطحی از ابهام در داده ها یا تناقض در داده ها باشد. ابهام یا تناقض داده‌ها باعث عدم قطعیت در داده های تولید شده توسط منبع می شود. به طور مثال یک سنسور اندازه‌گیری دما همواره با دقتی خاص تولید می‌شود. برای مثال، سنسور LM35 دارای دقتی معادل ۰,۱ درجه سانتی گراد می‌باشد. گاهی ممکن است همین سیستم تعیین دما، دچار نویز شده و دمای کاملا اشتباهی تولید نماید. [4]

می توان یکی از دستاورد های همجوشی داده ها را کاهش عدم قطعیت دانست (همه ی روش های پیشین توانایی برخورد با عدم قطعیت را ندارند.). ذکر این نکته در این بخش مفید است که همجوشی داده های حاصل از چند منبع غیر قطعی، ممکن است عملکرد پایین‌تری نسبت به تک تک منابع داشته باشد.[5]

۱–۳– همجوشی داده ها در شرایط عدم قطعیت

همجوشی داده ها در شرایط عدم قطعیت شامل دو بخش عمده می باشد؛ بخش اول همجوشی داده‌ها و بخش دیگر شناسایی و برخورد با عدم قطعیت و تناقضات می‌باشد. با توجه به تعاریف ارائه گذشته و بخش های مهم مذکور، در این پایان نامه تعریف جدید و کاملتری برای همجوشی داده ها ارائه شده است. همجوشی داده ها در شرایط عدم قطعیت عبارت است از ترکیب چند

منبع غیر قطعی برای ساختن یک بازنمایی مؤثر برای انسان یا ماشین با توانایی شناسایی و برخورد با عدم قطعیت و تناقضات.

۱–۴– هدف پایان نامه

مهمترین هدف این پایان نامه شناسایی و رسیدگی به عدم قطعیت و تناقضات داده ها در همجوشی برای سیستم های ناشناس می باشد. غالبا روش های پیشین همجوشی داده ها مانند فیلتر کالمن با مدل رفتاری سیستم کار می‌کنند؛ این بدین معنا است که برای همجوشی داده ها از یک جواب نظری استفاده می کنند. در سیستم هایی که مدل رفتاری آنها در دسترس ما نیست نمی‌توان از الگوریتمی مانند فیلتر کالمن استفاده نمود. باید به این موضوع توجه نمود که الگوریتم ارائه شده در این پایان نامه، می تواند برای سیستم های با مدل مشخص کار نماید اما بدیهی است با حذف مدل سیستم فرآیند همجوشی و رسیدگی به خطا ها مشکل تر خواهد شد. مهمترین تفاوتی که در سیستم هایی با مدل رفتاری ناشناس و شناسا وجود دارد این است که در سیستم های شناسا، عملیات همجوشی در شرایط عدم قطعیت حتی با یک سنسور هم امکان پذیر است زیرا مدل سیستم در تولید داده‌ی همجوشی شده، کمک خواهد کرد. در سیستم های ناشناس برای رسیدگی به تناقضات به تعداد زیادی سنسور نیاز خواهد بود. اگر منابع دارای عدم قطعیت باشند، ممکن است همجوشی داده های آنها از واقعیت دور بوده و ارزیابی سیستم از محیط اشتباه باشد. روش ارائه شده در این پایان نامه، این امکان را کاهش می‌دهد.

فصل دوم – ادبیات تحقیق و بررسی پژوهش های مرتبط

۲–۱–بررسی پژوهش های گذشته

در کارهای گذشته برای همجوشی داده ها سطوح متفاوتی ارائه شده است. همجوشی داده ها بر اساس سطح داده هایی که برای همجوشی مورد استفاده قرار می گیرند به چهار سطح پایین، متوسط، بالا و چند سطحی[1] تقسیم می شوند. سطح پایین که معمولا بر روی یک سیگنال فقط انجام می شود و با حذف نویز یک همجوشی ارائه می دهد. روش تبدیل موجک، یکی از روش های سطح پایین است. سطح متوسط شامل روش های ارائه یک بازنمایی از محیط می باشد. به طور مثال با یک شبکه عصبی می‌توان ورودی های متفاوتی مانند دما و سرعت باد را دریافت و یک خروجی از آن محاسبه نمود. روش های سطح بالا، شامل ارائه یک استنتاج از محیط هستند. فیلتر کالمن از همین دسته خواهد بود. بالاخره روش چند سطحی، ترکیبی از سه روش بالا خواهد بود.[6]

منابع در همجوشی داده ها می توانند مکمل یکدیگر باشند به طور مثال می توانند قسمت های مختلفی از محیط را ارزیابی نمایند. به طور مثال در اتاق ســـرور، دمای نقاط متفاوت یک اندازه نخواهد بود. منابع مســـتقل می توانند یک پدیده را اندازه گیری نمایند. این اطلاعات به منظور افزایش اطمینان و دقت با هم ترکیب می‌شـــوند. این همجوشـــی می تواند باعث کاهش حجم ترافیک نیز شود. همچنین منابع مستقل می توانند داده‌هایی را فراهم آورند که اطلاعات جدید و احتمالا پیچیده‌تری از این داده‌ها قابل اســتخراج باشد؛ به عنوان مثال ترکیب داده‌های فاصله و زاویه برای محاسبه‌ی مکان شی. [1]

تکنیک های زیادی برای همجوشی داده ها وجود دارد یکی از معروفترین و پرکاربرد ترین روش ها، فیلتر کالمن می باشد. فیلتر کالمن به خوبی تناقضات و عدم قطعیت را شناسایی و با آن

[1] Hybrid or multi-level

برخورد می‌کند. فیلتر کالمن برای همجوشی داده ها نیاز به داشتن مدل رفتاری سیستم دارد. [7] [8]

روش های دیگری مانند فیلتر H-infinity نیز برای همجوشی استفاده شده است[9] . در کارهای گذشته، روش‌های هوش مصنوعی نیز به کرّات مورد استفاده قرار گرفته است. منطق فازی[10] ، شبکه عصبی[11] و الگوریتم ژنتیک[12] از مواردی است که می توان به عنوان روش های هوش مصنوعی نام برد. روش تبدیل موجک، که روشی سطح پایین برای همجوشی داده ها می باشد بارها مورد استفاده قرار گرفته است.[14] [13]

در مقالاتی از روش های ترکیبی استفاده کرده اند؛ به طور مثال نویسندگان در مقاله [8] از ترکیب شبکه عصبی و فیلتر کالمن استفاده نموده اند. برای همجوشی داده ها در منابع از روش های فیلتر بیزین [15] و تئوری دمپستر-شفر[16] نیز استفاده شده است.

با مطالعه روش های پیشین تلاش ما بر آن بود تا با روش فیلتر کالمن همجوشی داده ها را انجام دهیم اما تلاش ناموفق ما سرانجامی نداشت زیرا این روش نیاز به مدل رفتاری سیستم دارد. در ادامه فیلتر کالمن، تئوری بیزین و تئوری دمپستر-شفر نیز توضیح داده شده است. روش ارائه شده در این پایان نامه، دارای ایده از فیلتر کالمن و روش بیزین است.

۲-۱-۱- تئوری همجوشی بیزین

تئوری بیزین یا تئوری باور بیزین یک گراف جهت‌دار است که مجموعه‌ای از متغیرهای تصادفی و نحوه ارتباط مستقل آنها را نشان می‌دهد. به عنوان نمونه یک شبکه بیزی می‌تواند نشان دهنده ارتباط بین علت بارش باران با میزان بارش آنها(سیل،زیاد،کم،بدون باران) باشد. بنابراین با داشتن

عوامل بارش می توان احتمال میزان بارش باران را پیش‌بینی نمود. تئوری بیزین، روشی برای برای شناسایی روابط احتمالی به منظور پیشگویی یا ارزیابی بین کلاس عضویت ها است.

تئوری همجوشی بیزین از روی روابط کلاسیک بیز بدست آمده است. تئوری بیزین و فیلتر کالمن دارای ایده های اولیه یکسانی هستند. قبل از توضیح بیشتر تئوری بیز، یکبار سناریوی مسئله آورده شده است.

فرض کنید اطلاعاتی از منابع مختلف برای یک سیستم وجود دارد و قرار است اطلاعات دیگری تخمین زده شود و یا با در کنار هم قرار دادن همین اطلاعات، یک بازنمایی از سیستم ساخته شود. به طور کلی می توان هدف را تخمین حالت[1] سیستم در نظر گرفت.

با توجه به سناریوی بالا فرض کنید یک سیستم گسسته در زمان با معادلات حالت و خروجی های زیر دارید. اگر فرض کنید در لحظه یا نوبت k هست x_k حالت فعلی، x_{k-1} حالت قبلی است. u_k ورودی سیستم در لحظه k ، w_k و v_k به ترتیب نویز فرآیند و نویز اندازه گیری در لحظه جاری هستند. y_k نیز خروجی سیستم در همین لحظه می باشد.

$$x_k = f_k(x_{k-1}, u_k, w_k) \qquad (۱-۲)$$

$$y_k = h_k(x_k, u_k, v_k) \qquad (۲-۲)$$

هدف همجوشی داده ها، مشخص نمودن حالت سیستم است. اگر فرض کنید بدانیم حالت اولیه سیستم کدام است و اندازه گیری ها را تا لحظه k داشته باشیم هدف محاسبه توزیع احتمالی

[1] State estimation

$p(x_k|Y_k)$ می باشد. با محاسبه این توزیع احتمالی، نقاطی حالت سیستم را مشخص می نمایند که احتمال مفروض بالاتر باشد.

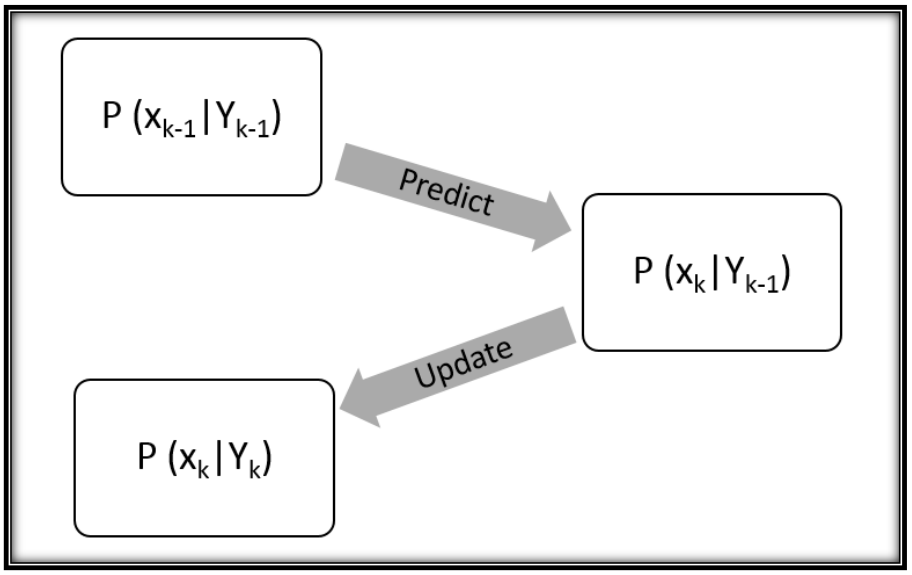
فرض کنید سیستم به شکل زیر تعریف شده است.

$$x_k = f_k(x_{k-1}, w_k) \quad (۳-۲)$$

$$y_k = h_k(x_k, v_k) \quad (۴-۲)$$

$$Y_k = \{y_0, y_1, \dots, y_k\} \quad (۵-۲)$$

اساس همجوشی بیزین در شکل ۲- ۱ آمده است. در این شکل مشاهده می نمایید که محاسبه توزیع احتمالی $p(x_k|Y_k)$ در دو مرحله انجام می شود. در مرحله اول بدون توجه به مشاهدات در لحظه k یک پیش بینی از سیستم ارائه می شود(همان $p(x_k|Y_{k-1})$). سپس با افزودن مشاهدات در لحظه k پیش بینی قبلی به روز رسانی می شود(همان $p(x_k|Y_k)$).



شکل ۲ - ۱ : اساس تئوری همجوشی بیزین

برای محاسبه پیش بینی و به روز رسانی از روابط زیر استفاده می‌شود.

$$p(x_k|Y_{k-1}) = \int p(x_k|Y_{k-1})p(x_k|Y_{k-1})dx_{k-1} \qquad (۶-۲)$$

$$\begin{aligned} p(x_k|Y_k) &= p(x_k|Y_k, y_k) \\ &= \frac{p(y_k|x_k)p(x_k|Y_{k-1})}{p(y_k|Y_{k-1})} \end{aligned} \qquad (۷-۲)$$

رابطه(۲-۷) برای به روز رسانی پیش بینی استفاده می شود. مخرج این کسر یک ضریب نرمال سازی است و می‌توان آن را از رابطه زیر محاسبه نمود.

$$p(y_k|Y_{k-1}) = \int p(y_k|x_k)p(x_k|Y_{k-1})dx_k \qquad (۸-۲)$$

اگر چه این روش، تئوری بسیار قوی ریاضیاتی دارد اما به صورت کلی حل روابط انتگرالی بالا دشوار است. بنابراین برای استفاده از آن برای همجوشی داده ها، باید سیستم مورد نظر کاملا مشخص باشد و مدل ریاضیاتی سیستم در اختیار ما باشد. بر مبنای همین تئوری، با تقریب آماری انتگرال های بالا، روش های دیگری برای همجوشی داده ها مانند particle filter ارائه شده است.[17]

۲-۱-۲- فیلتر کالمن

ایده اصلی فیلتر کالمن و فیلتر بیزین کاملا یکسان است و می توان بیان نمود فیلتر کالمن، حالت خاصی از تئوری همجوشی بیزین است. با توجه به مباحث مطرح شده برای تئوری همجوشی بیزین، کافی است شرایط خاص فیلتر کالمن را به تئوری بیزین اعمال نماییم. اگر فرض کنید سیستم خطی باشد و نویزهای سیستم، نویز نرمال باشند فرمول های تئوری همجوشی بیزین ساده شده و فیلتر کالمن ساخته می‌شود.

فرض کنید سیستم به صورت خطی و به شکل زیر تعریف شود. اگر فرض کنید در لحظه یا نوبت k هست x_{k+1} حالت بعدی، x_k حالت فعلی است. w_k و v_k به ترتیب نویز فرآیند و نویز اندازه گیری در لحظه جاری هستند. y_k نیز مشاهدات سیستم در همین لحظه می باشد.

$$x_{k+1} = F_k x_k + G_k w_k \quad (۹-۲)$$

$$y_k = H_k x_k + D_k v_k \quad (۱۰-۲)$$

$$Y_k = \{y_0, y_1, \dots, y_k\} \quad (۱۱-۲)$$

با توجه به آن که نویز های سیستم به شکل نرمال هستند، داریم:

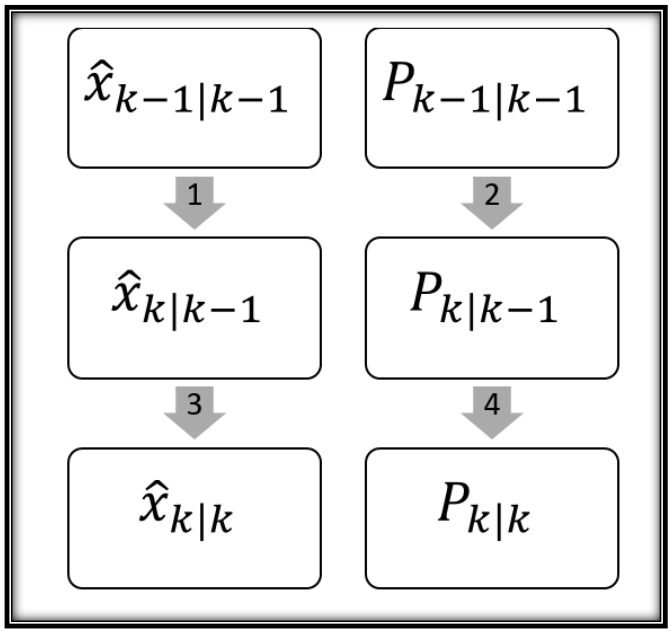
$$w_k \sim N(0, Q_k) \quad (۱۲-۲)$$

$$v_k \sim N(0, R_k) \quad (۱۳-۲)$$

$$p(x_k|Y_{k-1}) = N(\hat{x}_{k|k-1}, P_{k|k-1}) \quad (۱۴-۲)$$

$$p(x_k|Y_k) = N(\hat{x}_{k|k}, P_{k|k}) \quad (۱۵-۲)$$

برای محاسبه روابط ۱۴ و ۱۵ باید به ترتیب مقادیر کواریانس و میانگین ها را محاسبه نمود. این ترتیب در شکل ۲- ۲ نشان داده شده است.



شکل ۲- ۲: ترتیب پیش بینی ها در فیلتر کالمن

فرمول های محاسبه کواریانس و میانگین ها در زیر آمده است.

$$\hat{x}_{k|k-1} = F_{k-1}\hat{x}_{k-1|k-1} \qquad (۱۶-۲)$$

$$P_{k|k-1} = G_{k-1}Q_{k-1}G_{k-1}^{T} + F_{k-1}P_{k-1}F_{k-1}^{T} \qquad (۱۷-۲)$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - H_k\hat{x}_{k|k-1}) \qquad (۱۸-۲)$$

$$P_{k|k} = (1 - K_kH_k)\, P_{k|k-1} \qquad (۱۹-۲)$$

در روابط بالا فقط مقدار بهره کالمن مشخص نیست که از رابطه زیر محاسبه می شود.

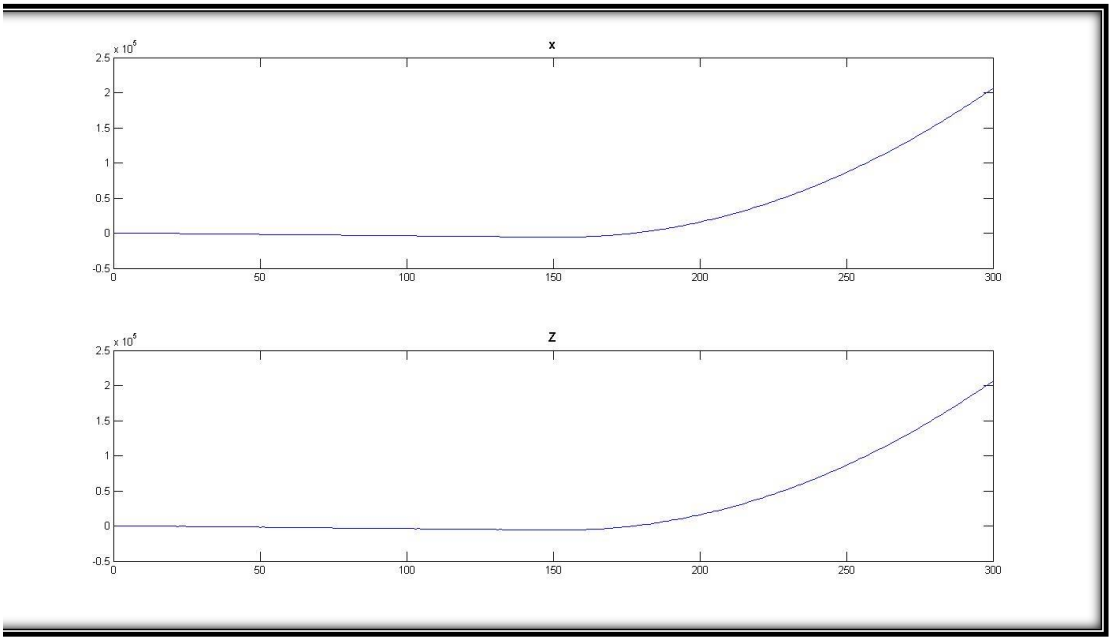
$$K_k = P_{k|k-1} H_k^T (H_k P_{k|k-1} H_k^T + D_k R_k D_k^T)^{-1} \qquad (۲۰-۲)$$

بنابراین با کمک روابط بالا می توان با فیلتر کالمن همجوشی را انجام داد اما مشکل اصلی در این بخش این است که معادلات و مدل ریاضیاتی سیستم در اختیار ما قرار ندارد. به عبارت دیگر، در صورتی که سیستم دارای مدل کاملی باشد فیلتر کالمن به خوبی همجوشی را انجام می دهد. شاید به نظر برسد که وقتی مدل رفتاری سیستم مشخص است، استفاده از همجوشی داده ها یا فیلتر کالمن چه فایده‌ای دارد. به سناریوی زیر توجه نمایید.

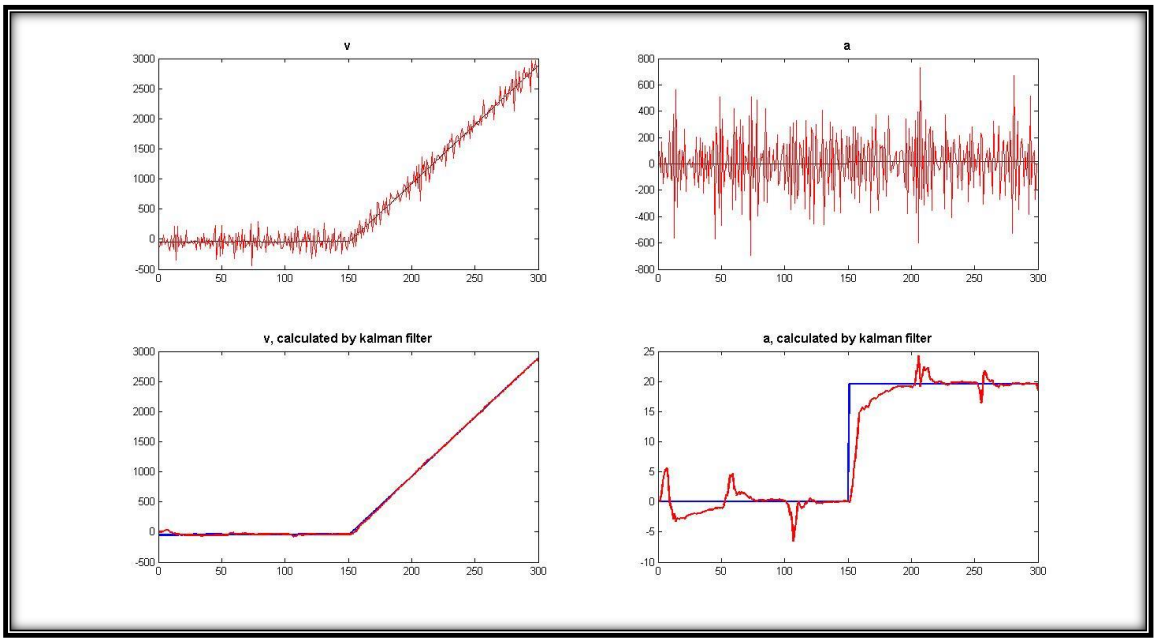
فرض کنید بخواهید رفتار یک "شیء متحرک" را پیش بینی نمایید. به طور مثال فرض کنید میزان جا به جایی شیء را دارید و می‌خواهید سرعت لحظه ای و شتاب لحظه ای شیء را محاسبه نمایید. اگر سیستم بدون نویز اندازه گیری شود، با مشتق از جا به جایی به سرعت لحظه ای دست می یابید. با دو بار مشتق از جا به جایی می توانید شتاب لحظه ای را محاسبه نمایید. لازم است یکبار دیگر تاکید شود که این روش مشتق گرفتن فقط و فقط برای زمانی پاسخگو است که سیستم بدون نویز مشاهده و اندازه گیری شود.

عمدتا نویز ماهیتی فرکانس بالا دارد و در واقع مشتق نویز مقدار زیادی دارد زیرا معمولا نویز ها دارای تغییرات زیادی هستند. بنابراین با دو بار مشتق گرفتن، شتاب لحظه ای سیستم محاسبه نمی‌شود. آزمایشات نشان می دهد که شتاب لحظه ای محاسبه شده به روش مشتق برای سیستم های دارای نویز، حتی شباهتی به شتاب لحظه ای واقعی ندارد و کاملا به شکل نویز خود را نشان می دهد. در این حالت، یکی از راه های گذر از این مشکل، استفاده از فیلتر کالمن است.

سناریوی بالا به کمک فیلتر کالمن پیاده سازی شده است. مقدار واقعی جا به جایی و مقدار اندازه گیری شده آن در شکل ۲- ۳ آمده است. متوسط خطای مطلق در مقدار اندازه گیری شده ۸۰,۰۹ می باشد.



شکل ۲- ۳: یک مثال برای فیلتر کالمن؛ بالا) مقدار واقعی جا به جایی، پایین) مقدار اندازه گیری شده جا به جایی



شکل ۲ - ٤:محاسبه سرعت و شتاب لحظه ای؛ بالا) محاسبه شده با روش مشتق، پایین) محاسبه شده با روش فیلتر کالمن

شکل ۲- ۴، سرعت لحظه ای و شتاب لحظه ای - محاسبه شده توسط دو روش مشتق و فیلتر کالمن را نشان می دهد. در نمودار های شکل ۲- ۴ ، مقادیر محاسبه شده با رنگ قرمز و مقادیر اصلی با رنگ آبی نمایش داده شده اند (مقدار واقعی شتاب و سرعت با مشتق از مقادیر بدون نویز جا به جایی محاسبه شده اند.). نمودار های بالایی، سرعت و شتاب لحظه ای را با روش مشتق نمایش می دهند. همان طور که مشاهده می نمایید نمودار های بالایی به جای شدت دارای نویز هستند و مقدار شتاب، کلا اشتباه است. در نمودار های پایینی، با کمک فیلتر کالمن همجوشی انجام شده است که مقدار سرعت به خوبی محاسبه شده است. اگر چه شتاب لحظه ای دارای خطا می باشد اما می تواند اطلاعات بیشتری به ما بدهد. با این مثال، مشخص می شود که اگر چه مدل ریاضیاتی سیستم در اختیار ما قرار دارد اما همیشه به خودی خود کارآمد نیست.

۲-۱-۳- تئوری دمپستر - شفر[1]

تئوری دمپستر-شفر یک تئوری استخراج شده از تئوری بیزین است. DST بر سه مفهوم فرضیات[2]، تکه‌های باور[3] و منبع داده[4] استوار است. فرضیات شامل تمام حالت های ممکن برای سیستم است و باور ها تمام زیر مجموعه های فرضیات است. به عبارت دیگر باور ها، مجموعه توانی فرضیات هستند. این از تفاوت های تئوری بیزین و تئوری دمپستر-شفر است. زیرا مجموعه هدف در تئوری بیزین فرضیات است ولی مجموعه هدف در DST مجموعه توانی فرضیات یا همان تکه های باور است. منبع داده شامل هر چیزی است که در کسب اطلاعات سیستم به ما کمک می کند.

برای همجوشی داده ها با این تئوری، از تابع mass استفاده می شود. تابع mass یک نگاشت از فضای تکه‌های باور به اعداد حقیقی می دهد.

برای محاسبه همجوشی دو حسگر از رابطه زیر استفاده می شود.

$$m^{1,2}(Z) = \frac{\sum_{A\cap B=Z} m^1(A) * m^2(B)}{\sum_{A\cap B=\emptyset} m^1(A) * m^2(B)} \qquad (۲۱-۲)$$

تعمیم رابطه بالا برای ۳ و N حسگر به شکل زیر خواهد بود.

[1] Dempster Shafer Theory (DST)
[2] hypotheses
[3] Piece of evidence
[4] Data source

$$m^{1,2,3}(Z) = \frac{\sum_{A\cap B\cap C=Z} m^1(A) * m^2(B) * m^3(C)}{\sum_{A\cap B\cap C=\emptyset} m^1(A) * m^2(B) * m^3(C)} \qquad (۲۲-۲)$$

$$m^{1,2,\dots,N}(Z) = \frac{\sum_{\cap A_i=Z}(\prod_{A_i} m^i(A_i))}{\sum_{\cap A_i=\emptyset}(\prod_{A_i} m^i(A_i))} \qquad (۲۳-۲)$$

هر آنچه مقدار تابع mass بیشتر باشد، احتمال آن باور بیشتر است. نکته مهم برای همجوشی داده‌ها به روش

دمپستر-شفر این است که مقادیر اولیه mass باید به کمک دینامیک سیستم، فرد خبره یا مدل ریاضی سیستم مشخص شود.

۲-۲- ادبیات تحقیق

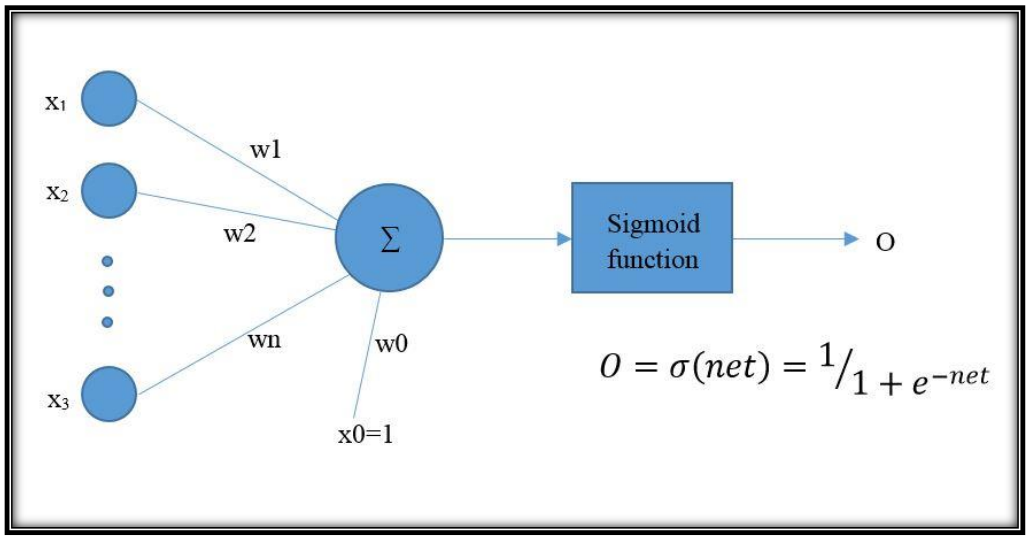
این بخش، به بررسی پایه و اساس روش های مورد استفاده در این پایان نامه پرداخته شده است. در فصل بعدی روش پیشنهاد شده، به طور مفصل توضیح داده شده است. در این بخش مفاهیمی مطرح شده است که برای درک روش ارائه شده در پایان نامه لازم است. همانطور که بعدا خواهید دید روش ارائه شده، از سه بخش خوشه بندی، پیش بینی گر MLP[1] و قانون به روز رسانی پیش بینی تشکیل شده است. در بخش (۲-۲-۱) مروری اجمالی بر شبکه عصبی MLP شده است. بخش (۲-۲-۲) شامل مفاهیم اصلی خوشه بندی است. در بخش بعدی ساختار دو مرحله ای فیلتر کالمن مورد بررسی قرار گرفته است که با الهام از آن روش ارائه شده ساخته شده است.

[1] Multi-Layer Perceptron

نهایتا در بخش (۲-۲-۴) شبکه حسگر بیسیم[1] توضیح داده شده است که می تواند خود سناریویی برای الگوریتم ارائه شده در این پایان نامه باشد.

۲-۲-۱- شبکه عصبی MLP

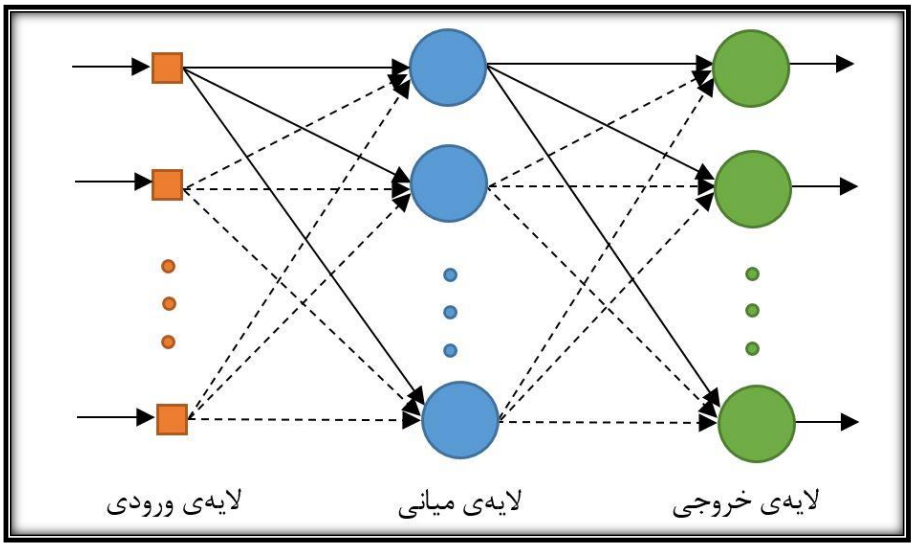
شبکه عصبی از تعداد دلخواهی نرون تشکیل می‌شود که نگاشتی از مجموعه ورودی را به خروجی انجام می‌دهد. شبکه عصبی MLP، متشکل از چند پرسپترون است. پرسپترون نوعی از شبکه عصبی برمبنای یک واحد محاسباتی است. یک پرسپترون برداری از ورودی‌ها را گرفته و یک ترکیب خطی از این ورودی‌ها را به عنوان خروجی ارائه می نماید. برای اینکه بتوانیم فضای تصمیم گیری را بصورت غیر خطی از هم جدا نماییم، لازم است تا هر سلول واحد را بصورت یک تابع غیر خطی تعریف نماییم. یک پرسپترون با تابع غیر خطی سیگموید در شکل ۲- ۵ آورده شده است.



شکل ۲- ۵: نمایش یک پرسپترون با تابع غیر خطی سیگموید

[1] Wireless Sensor Network (WSN)

MLPها که از کنار هم قرار دادن پرسپترون ها ایجاد شده اند، بر خلاف پرسپترون ها توانایی یادگیری مسائل با تصمیم گیری های متعدد را دارند.ساختار کلی شبکه عصبی MLP با یک لایه مخفی در شکل ۲- ۶ نشان داده شده است.



شکل ۲ - ٦:شاختار کلی شبکه MLP با یک لایه مخفی

برای آموزش شبکه MLP می توان از روش Back Propagation استفاده نمود. در این روش با محاسبه‌ی خطا و به روز رسانی وزن‌ها و دیگر مقادیر، در هر گام به بهبود شبکه پرداخته می شود. الگوریتم آموزش شبکه عصبی MLP در ادامه آمده است. در ابتدا باید شبکه ای با تعداد مشخص گره ورودی، گره مخفی و گره خروجی ایجاد کرد. سپس به همه وزن ها یک مقدار تصادفی کوچک اختصاص داده می شود. تا رسیدن به شرط پایانی (مثلا کوچک شدن خطا) مراحل زیر را انجام می پذیرد.

برای هر نمونه s متعلق به مثالهای آموزشی:

۱. مثال s به سمت جلو در شبکه انتشار داده می شود.

۲. خطای δ به سمت عقب در شبکه انتشار داده می شود.

که برای انتشار مثال s مقدار خروجی هر نرون محاسبه می شود تا خروجی نرون های لایه آخر مشخص گردند. همچنین خطا به صورت زیر محاسبه می شود.

$$\begin{cases} \delta_j = o_j(1-o_j)(t_j - o_j) & j\ is\ output\ node \\ \delta_j = o_j(1-o_j)\sum_k \delta_k w_{kj} & O.W. \end{cases} \qquad (۲۴-۲)$$

که در آن o_j خروجی نرون j، t_j خروجی مورد انتظار مربوط به نرون j می باشد. نهایتا مقادیر وزن ها با رابطه زیر به روز رسانی می شود که در آن λ ضریب یادگیری شبکه عصبی است.

$$w_{ji} = w_{ji} + \lambda \delta_j o_i \qquad (۲۵-۲)$$

ضریب یادگیری پارامتری است که با استفاده از آن به اصلاح وزن‌ها پرداخته می شود. هرچه ضریب یادگیری بیشتر باشد، میزان تغییرات وزن ها بیشتر خواهد بود و بالعکس هرچه ضریب یادگیری کم تر باشد میزان تغییرات وزن ها نیز کاهش می یابد. افزایش ضریب یادگیری و در نتیجه افزایش تغییرات وزن ها باعث می شود در مواردی که وزن قبلی فاصله زیادی از مقدار

اصلیش دارد با سرعت بیشتری به مقدار واقعی آن نزدیک شود. این امر کمک می کند که شبکه عصبی دچار مینیمم محلی نشود. در پایان یادگیری شبکه عصبی، بهتر است ضریب یادگیری کمتر باشد تا میزان تغییرات ناگهانی کاهش یابد. بنابراین پیشنهاد می-شود در ابتدا برای اجتناب از جواب های محلی مقدار ضریب یادگیری کمتر بوده و با پیشرفت یادگیری مقدار این ضریب هربار کمتر شود.

۲-۲-۲- خوشه بندی داده ها

در این از پایان نامه، در ابتدا می خواهیم به بیان سه مفهوم دسته بندی[۱]، خوشه بندی[۲] و افراز یک مجموعه[۳] بپردازیم. در حالت کلی، افراز یک مجموعه به معنای انتخاب بعضی از زیر مجموعه های یک مجموعه به قسمی که اشتراک هر یک از زیر مجموعه ها با یکدیگر تهی و اجتماع همه آنها با یکدیگر برابر با مجموعه اصلی باشد. [18]

خوشه بندی، به تقسیم یک مجموعه به چند خوشه گفته می شود که اعضای کلاستر به یکدیگر شبیه بوده و اعضای کلاستر های متفاوت شباهت کمی با هم داشته باشند. یا به عبارت دیگر، فاصله اعضای یک کلاستر از یکدیگر کمتر از فاصله بین نقاط غیر هم کلاستر باشد.[19] در خوشه بندی، کلاستر ها در روند کلاسترینگ تشکیل شده و بنابراین هیچ نام[۴] یا صفت خاصه ای برای آنها در طول فرآیند تعریف نمی شود. پس از انجام کلاسترینگ نیاز به یک سیستم خبره یا فرد خبره ای وجود دارد که برای هر خوشه، یک نام در نظر بگیرد یا آنها را تحلیل نماید.[20]

[1] classification
[2] clustering
[3] Partition of a set
[4] label

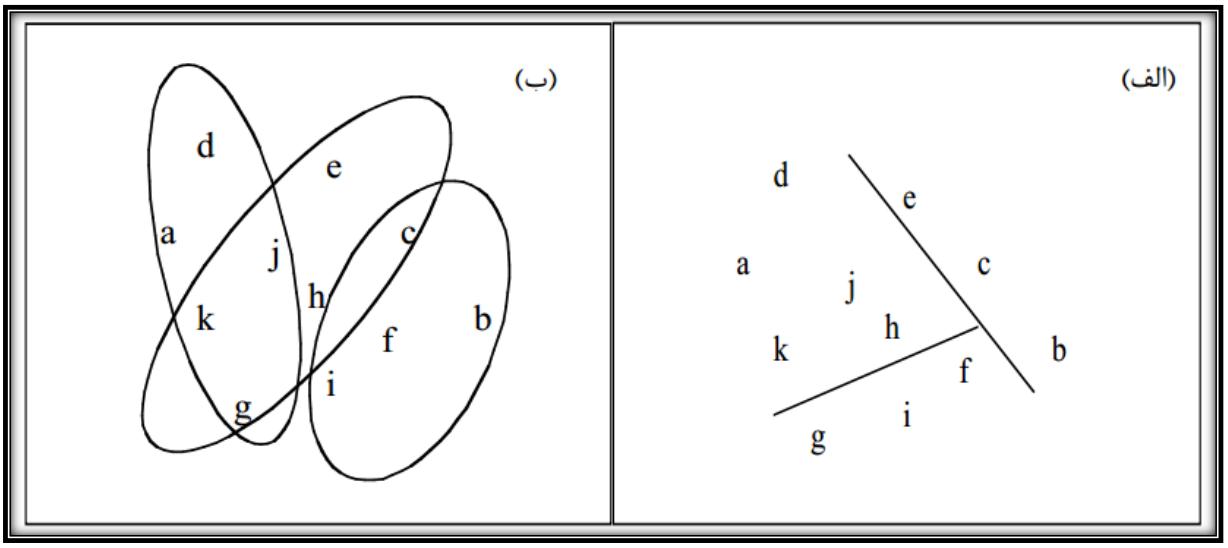
دسته بندی به تقسیم یک مجموعه به چند دسته گفته می شود که قبل از عمل دسته بندی، تعداد دسته ها، ویژگی های کلاس ها و صفات خاصه آنها تعریف شده است.هدف اصلی دسته بندی، قرار دادن نمونه در دسته‌های از قبل تعیین شده است.[21]

با توجه به تعاریف ارائه شده، در صورتی که فقط روش های کلاسیک را مدنظر قرار دهیم، هر کلاسترینگ و دسته بندی نوعی افراز مجموعه است. افراز مجموعه یک مفهوم ریاضیاتی عام است که شالوده و بنیان خوشه بندی و دسته بندی را تشکیل می دهد. [18]

روش های زیادی برای خوشه بندی، دسته بندی و افراز یک مجموعه وجود دارد، به علت شباهت بیشتر روش پیشنهاد شده در این پایان نامه با خوشه بندی، در باب خوشه بندی توضیحات بیشتری گفته خواهد شد.

تکنیک‌های خوشه بندی را به دو دسته خوشه بندی انحصاری[1] و خوشه بندی غیر انحصاری[2] تقسیم می‌نمایند. روش‌های انحصاری روش هایی هستند که هر نمونه، فقط و فقط در یک گروه جای می‌گیرد و در حالی که در روش‌های غیر انحصاری یک عضو می تواند متعلق به چند دسته باشد شکل ۲- ۷) . روش های کلاسیک خوشه بندی از جمله روش های انحصاری و روش های فازی از جمله روش های غیر انحصاری هستند.[22]

[1] exclusive
[2] non-exclusive



شکل ۲ - ۷:روش های خوشه بندی، الف) انحصاری؛ ب)غیر انحصاری

تکنیک های خوشه بندی را می توان به دو دسته روش های تکرار توالی و روش های همزمان تقسیم کرد. روش های همزمان، با یکبار اختصاص نمونه ها به خوشه ها عملیات خوشه بندی به اتمام می‌رسد. در حالی که در روش های تکرار توالی با چند بار محاسبه بر روی نمونه و تغییر خوشه های آنها انجام می شود. [20]

روش های سلسله مراتبی[1] و روش های غیر سلسله مراتبی[2] نیز تقسیم بندی دیگری برای خوشه بندی می‌باشد. در روش‌های سلسله مراتبی، خوشه ها هر کدام به خوشه های ریز تری تقسیم شده و خود نیز عضو خوشه بزرگتری خواهد بود. روش های غیر سلسله مراتبی، خوشه ها مستقل از یکدیگر هستند. خود روش های سلسله مراتبی نیز به دو دسته بالا به پایین و پایین به بالا تقسیم می‌شوند. [20]

[1] hierarchical approaches
[2] Non- hierarchical approaches

۲-۲-۲-۱- الگوریتم K-means

در کتاب [23]، روش K-means با توضیحات کامل آورده شده است. روش K-means با آن که روشی ساده می باشد اما پایه و اساس بسیاری از روش های خوشه بندی مانند خوشه بندی فازی می باشد. علت آن که این روش در این بخش توضیح داده شده است این است که می‌تواند به عنوان پایه ای برای روش ما نیز باشد. این الگوریتم، الگوریتمی انحصاری، تکرار متوالی و غیر سلسله مراتبی می‌باشد. برای این الگوریتم شکلهای مختلفی بیان شده است. در همه آن ها تعداد خوشه ها مشخص و ثابت است و موارد زیر را محاسبه می‌نمایند.

۱. بدست آوردن نقاطی به عنوان مراکز خوشه‌ها
۲. نسبت دادن هر نمونه داده به یک خوشه
۳. محاسبه دوباره مراکز خوشه ها

در واقع،می توان خوشه بندی با الگوریتم k-means را معادل با بهینه سازی تابع هدف قرار داد. فرض کنید تابع هدف زیر برای الگوریتم k-means تعریف شده است.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n_j} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (۲۶-۲)$$

که در آن c_j مرکز خوشه j ام، k تعداد کل خوشه ها و n_j تعداد اعضای خوشه j ام می باشد. همچنین $x_i^{(j)}$ داده i ام در خوشه j ام و نیز || || معیار فاصله بین نقاط می باشد. مراحل پیاده سازی الگوریتم k-means در زیر آورده شده است. توجه کنید که تعداد کل خوشه ها باید مشخص باشد و آن را با k نشان می‌دهیم.

۱. انتخاب k مرکز خوشه به صورت تصادفی

۲. هر نمونه داده به خوشه‌ای تعلق می گیرد که فاصله داده مذکور تا آن خوشه کمتر از فاصله داده تا سایر مراکز خوشه ها باشد.

۳. مراکز خوشه ها دوباره محاسبه می شود. (مرکز هر خوشه عبارت است از میانگین داده های متعلق به هر خوشه)

مراحل ۲ و ۳ تا زمانی که دیگر هیچ تغییری در مراکز خوشه‌ها حاصل نشود، ادامه می یابد.

۲-۲-۲-۲- الگوریتم Single-Link

با این روش یکی از قدیمی‌ترین و ساده‌ترین روش های خوشه‌بندی است و شباهت بسیار زیادی با روش استفاده شده در این پایان نامه را دارد. این روش یا همان تکنیک نزدیک ترین همسایه[1]، جزء روشهای خوشه‌بندی سلسله مراتبی و انحصاری محسوب می‌شود[24] . در این روش معیار شباهت بین دو خوشه ci و cj از رابطه زیر محاسبه می‌شود.

$$d_{c_i c_j} = \min_{x_k \in c_i,\ x_l \in c_j} \{\|x_k - x_l\|\} \qquad (۲۷-۲)$$

[1] Nearest Neighbor

در واقع در این روش شباهت بین دو خوشه، کمترین فاصله بین یک عضو از یک خوشه با عضوی از خوشه دیگر است. این همان مفهومی است که ما در فصل های بعدی برای فاصله بین خوشه ها تعریف خواهیم کرد.

برای پیاده سازی این خوشه بندی، الگوریتم زیر را می توان استفاده نمود.

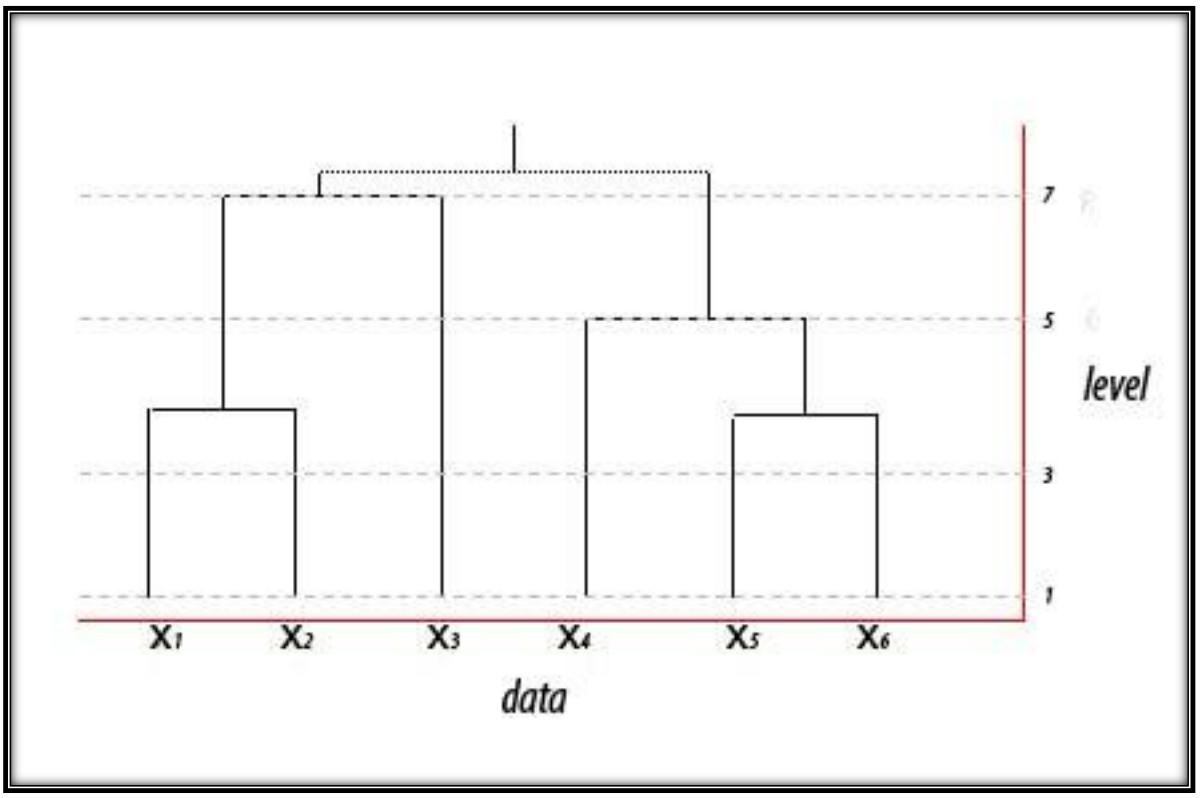
۱. هر داده به عنوان یک خوشه در نظر گرفته می‌شود.
۲. فاصله بین خوشه ها محاسبه می شوند.
۳. خوشه هایی که کمترین فاصله با یکدیگر را دارند با هم ترکیب و یک خوشه جدید می‌سازند. اگر چند گروه از خوشه ها، دارای فاصله یکسان از همدیگر و برابر با کمترین فاصله را داشتند، هر گروه به تنهایی ترکیب شده و یک خوشه را می سازد.

مراحل ۲ و ۳ آنقدر تکرار می شود تا همه داده ها در یک خوشه قرار بگیرند. هر بار که مراحل ۲ و ۳ تکرار می‌شوند یک خوشه بندی جدید ارائه می‌شود بنابراین می توان گفت که این خوشه بندی دارای سطوح متفاوت است. در شکل ۲- ۸ مثالی از خوشه بندی برای داده های موجود در جدول۲- ۱، آورده شده است. توجه داشته باشید به طور مثال در سطح ۵ ، سه خوشه وجود دارد که $\{x_4,x_5,x_6\}$، $\{x_3\}$ و $\{x_1,x_2\}$ می باشد.

جدول۲- ۱:مثال برای خوشه بندی با روش single-link – فاصله بین داده ها

فاصله	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	8	26	48	24	16
x_2	8	0	20	44	22	20

x_3	26	20	0	14	6	18
x_4	48	44	14	0	12	36
x_5	24	22	6	12	0	17
x_6	16	20	18	36	17	0



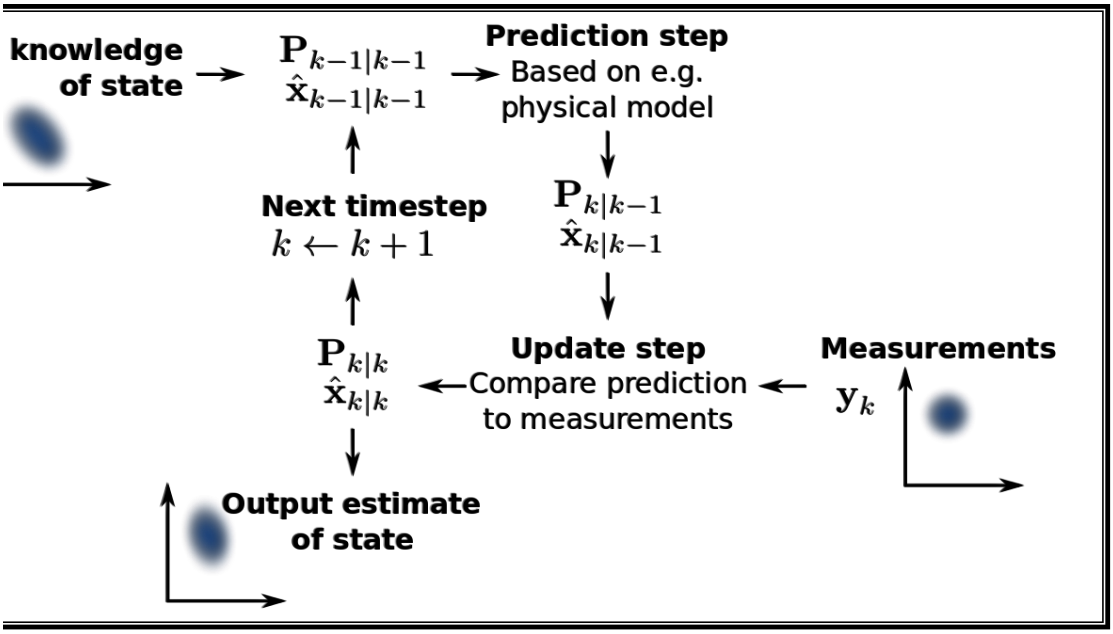
شکل ۲ - ۸:مثالی از روش خوشه بندی single-link

۲-۲-۳- ساختار کلی فیلتر کالمن

فیلتر کالمن، یکی از معروفترین روش های همجوشی داده ها می باشد. اگرچه در این پایان نامه، از فیلتر کالمن یا همجوشی بیزین استفاده نشده است اما شاید بتوان ساختار کلی آنها را با ساختار

کلی روش ارائه شده در این پایان نامه شباهت داد. در فیلتر کالمن، تخمین حالت[1] بر اساس پیش بینی[2] به کمک نتایج همجوشی های گذشته و به روز رسانی[3] پیش بینی انجام می‌شود؛ این وجه اشتراک مهمی بین روش ارائه شده و فیلتر کالمن می‌باشد.

فیلتر کالمن، توانایی همجوشی منابع در شرایط عدم قطعیت را دارد. فیلتر کالمن در هر مرحله، یک پیش‌بینی از حالت سیستم ارائه می دهد. سپس پیش بینی ارائه شده توسط اندازه گیری ها و مشاهدات منابع در لحظه جاری و همچنین مدل انتقال حالت ها[4]، به روز رسانی می شود.(شکل ۲- ۹) فیلتر کالمن برای پیش بینی حالت سیستم، از مشاهدات و اندازه‌گیری ها در لحظه جاری استفاده نمی نماید.[8]



شکل ۲ - ۹:ساختار کلی فیلتر کالمن

[1] state
[2] prediction
[3] update
[4] state transition model

اگرچه فیلتر کالمن در کارهای گذشته زیادی وجود داشته، اما از آنجایی که در سیستم های بدون مدل رفتاری در دسترس، کارآمد نیست؛ در این پایان نامه با الهام گرفتن از بنیان این روش، روش دیگری ارائه شده است.

۲-۲-۴- شبکه حسگر بیسیم[1]

شبکه های حسگر بیسیم، مجموعه ای از حسگر ها هستند که در محیط پخش می شوند و با مشاهده و اندازه‌گیری پدیده ها، یک بازنمایی از محیط ارائه می کنند. این حسگر ها به همراه سایر تجهیزات الکترونیکی تشکیل یک منبع را داده و منابع با ارسال مشاهدات خود به سرور، فعالیت خود را به پایان می رسانند. سرور وظیفه همجوشی این مشاهدات را دارد. شایان ذکر است در کارهای گذشته در مواردی همجوشی مشاهدات در سطح منابع هم وجود دارد. منابع موجود در WSN ها، معمولا از باتری استفاده می کنند؛ یکی از مهمترین مسائل این شبکه ها، کنترل مصرف انرژی می باشد[25] .

اگرچه آزمایشات روش ارائه شده در این پایان نامه، بر روی داده های حقیقی هواشناسی انجام شده است، اما خوب است سناریوی دیگری را طراحی کنیم. پیاده سازی این الگوریتم برای این سناریو می‌تواند در کارهای آتی مورد بررسی قرار بگیرد.

دقت کنید اصطلاحات منابع و سنسور ها به جای یکدیگر استفاده می شوند؛ تصور کنید منابع در جنگل برای شناسایی آتش سوزی، پراکنده گردیده است. همانطور که قبلا ذکر شد، بیشتر این منابع از باتری استفاده می نمایند. ارسال دائم مشاهدات باعث مصرف انرژی و عمر کوتاه

[1] Wireless Sensor Network(WSN)

باتری می شود. هدف این سناریو، شناسایی آتش سوزی بوده است یا به صورت دقیق تر هدف "شناسایی آتش سوزی با دریافت حداقل مشاهدات سنسورها" است.

فصل سوم – روش ارائه شده در پایان نامه و نتایج تجربی

۳-۱-طراحی سناریو

سیستم مورد نظر در این پروژه دارای دو ویژگی عمده می باشد؛ اولا سنسور ها دارای عدم قطعیت و تناقض هستند، ثانیا مدل رفتاری سیستم در اختیار ما قرار ندارد که بتوان داده های متناقض را حذف نمود. نداشتن مدل رفتاری سیستم، به معنای در اختیار نداشتن روابط و معادلات ریاضی برای سیستم است؛ به عبارت دیگر، ممکن است سیستمی دارای معادلات ریاضی باشد ولی ما آن‌ها را در اختیار نداشته باشیم. این دو ویژگی ما را به این سمت سوق می دهد که تعداد سنسور ها افزایش یابد. در بخش نتیجه گیری در مورد تعداد سنسورها بحث خواهد شد. دقت کنید منابع موجود برای سیستم، باید با یکاهای یکسانی اندازه گیری خود را انجام دهند به طور مثال، با الگوریتم ارائه شده در این پایان‌نامه، نمی توان داده های مربوط به سرعت باد و دمای هوا را همزمان ترکیب نمود.

قبل از بیان الگوریتم، سناریوی مسئله را به صورت دقیق مطرح شده است.

فرض کنید چند منبع یک پدیده را مشاهده یا اندازه گیری نمایند. هدف، ارائه یک بازنمایی[1] برای ترکیب مشاهدات منابع می باشد. بازنمایی ارائه شده در این پایان نامه، باید برای تصمیم گیری انسان یا ماشین کارآمد باشد. باید دو فرض اساسی را برای سناریو در نظر گرفت؛ اولا منابع موجود دارای سطحی از عدم قطعیت و تناقض هستند، ثانیا مدل رفتاری و روابط ریاضیاتی سیستم در اختیار ما قرار ندارد. همه منابع یک پدیده را با واحد اندازه گیری یکسانی مشاهده می نمایند و بر روی مشاهدات آنها عملیات محاسبه فاصله تعریف شده است.

حال با توجه سناریوی بالا، یک الگوریتم کارآ برای حل این سناریو طراحی شده است.

[1] Representation

۳–۲– الگوریتم برخورد با همجوشی داده ها برای سناریوی مفروض

در این بخش از پایان نامه، به معرفی الگوریتم پیشنهاد شده برای همجوشی داده ها پرداخته شده است. در سناریوی بالا، فرض کنید m منبع غیر قطعی در لحظه t مجموعه مشاهدات $Z_t=\{z_{1t}, z_{2t}, ...,z_{it},..., z_{mt}\}$ را دارند که در آن z_{it} مشاهده یا اندازه گیری منبع i ام در لحظه t باشد. بنابراین الگوریتم زیر برای این منظور ارائه شده است.

۱. خوشه بندی[1] مجموعه Z_t

۲. در صورتی که یکی از دسته ها شامل بیش از نیمی از سنسور ها بود، مرکز دسته به عنوان جواب ارائه می شود و الگوریتم به پایان می رسد.

۳. درصورتی که داده ها به چند دسته بزرگ تقریبا هم اندازه تقسیم شدند فقط داده های این چند دسته برای همجوشی استفاده می شوند و بقیه حذف خواهند شد.

۴. پیش بینی داده برای لحظه جاری از روی همجوشی های گذشته با کمک شبکه عصبی MLP

۵. به روز رسانی پیش بینی با کمک الگوریتم ارائه شده در بخش بعدی

۳–۲–۱– الگوریتم خوشه بندی طراحی شده

در این روش، لازم است مجموعه Z_t به خوشه هایی تقسیم شود که اعضای خوشه ها به یکدیگر نزدیک باشند. برای خوشه بندی، یک روش ارائه شده است که محصول آن یک خوشه بندی یکتا است. در این پایان نامه، برای به دست آوردن این خوشه‌بندی یکتا، الگوریتم زیر پیشنهاد شده است.

[1] Clustering

فرض کنید قرار است $Z_t=\{z_{1t}, z_{2t}, ..., z_{mt}\}$ با فاصله d خوشه بندی شود. برای هر داده از بردار Z_t عملیات زیر انجام می‌شود.

۱. اگر خوشه ای وجود دارد که فاصله آن با داده جدید کوچکتر یا مساوی فاصله d باشد، داده مورد نظر به خوشه افزوده می‌شود در غیر اینصورت یک خوشه با داده مورد نظر تشکیل می‌گردد.

۲. فاصله بین هر دو خوشه بررسی می شود در صورتی که دو خوشه با فاصله کوچکتر یا مساوی d وجود داشت، آنها ادغام می‌شوند.

در این پایان نامه منظور از فاصله یک داده تا یک خوشه، کمترین فاصله داده مورد نظر با تک تک عناصر خوشه می‌باشد. همچنین منظور از فاصله دو خوشه، کمترین فاصله بین هر داده از یک خوشه با خوشه دیگر است. در طراحی سناریو فرض شده است که عملیات محاسبه فاصله بین دو مشاهده امکان پذیر و مشخص است.

در حالت کلی، افراز یک مجموعه به معنای انتخاب بعضی از زیر مجموعه های یک مجموعه به قسمی که اشتراک هر یک از زیر مجموعه ها با یکدیگر تهی و اجتماع همه آنها با یکدیگر برابر با مجموعه اصلی باشد.[18]

فرض کنید $Z_t=\{z_{1t}, z_{2t}, ..., z_{mt}\}$ مجموعه مشاهدات سنسورها در لحظه t و m تعداد منابع یا سنسور ها و d فاصله افراز ها باشد. بنابراین می‌توان خوشه بندی مطرح شده را یکی از افراز های موجود از مجموعه بالا در نظر گرفت به شرطی که:

۱. اجتماع همه زیر مجموعه های منتخب، برابر با Z_t باشد.

۲. فاصله بین هر دو زیر مجموعه بیشتر از d باشد.

۳. z_{it} و z_{jt} عضو یک زیر مجموعه هستند اگر و تنها اگر فاصله بین آنها کوچکتر یا مساوی d باشد.

شرایط بالا، کفایت می کند که افراز مجموعه Z_t به صورت یکتا انجام شود. همانطور که قبلا ذکر شد در صورتی که یکی از دسته ها شامل بیش از نیمی از سنسور ها باشد، انتظار داریم داده های همین زیرمجموعه برای بازنمایی همجوشی منابع کافی باشد و احتمالا بقیه منابع دارای میزانی خطا هستند. دقت کنید اگر تعداد منابع کم باشد، به طور مثال ۳ منبع، این قانون دارای خطای زیادی است و انتظار برای همجوشی صحیح کاهش می یابد. بنابراین اینجا محدودیتی برای تعداد منابع وجود دارد. اگر در این مرحله، دسته ای شرایط مورد نظر را داشت(یا دقیقا یک خوشه با شرایط قبلی وجود دارد یا کلا وجود ندارد) میانگین دسته به عنوان جواب ارائه می‌شود و الگوریتم به پایان می رسد.

درصورتی که مشاهدات به چند خوشه تقریبا بزرگ و تعداد کم خوشه کوچک تقسیم شدند فقط مشاهدات چند خوشه بزرگ برای همجوشی استفاده می شوند و بقیه حذف خواهند شد. در اینجا انتظار می‌رود که خوشه‌های کوچک شامل میزانی عدم قطعیت و تناقض هستند و می‌توانند عملکرد کلی سیستم را کاهش دهند.

۳–۲–۲– پیش بینی گر Multi-Layer Perceptron

MLP یکی از انواع شبکه های عصبی با توانایی نگاشت ورودی به خروجی می باشد. در این پایان نامه برای پیش‌بینی نتیجه همجوشی، از یک شبکه عصبی MLP با یک لایه مخفی استفاده شده است. لایه مخفی در پیش‌بینی‌گر مذکور و لایه ورودی دارای تعداد مساوی از نرون می باشند.

برای تولید ورودی پیش‌بینی‌گر، از همجوشی های مراحل قبلی استفاده می شود. فرض کنید به نتایج همجوشی منابع در زمان های گذشته، ورودی های خام برای شبکه عصبی گفته می‌شود. به عبارت دیگر، اگر بخواهید همجوشی داده ها در لحظه t را محاسبه نمایید همجوشی داده ها تا لحظه t-1 به عنوان ورودی های خام در نظر گرفته می‌شود. برای ساختن ورودی شبکه عصبی

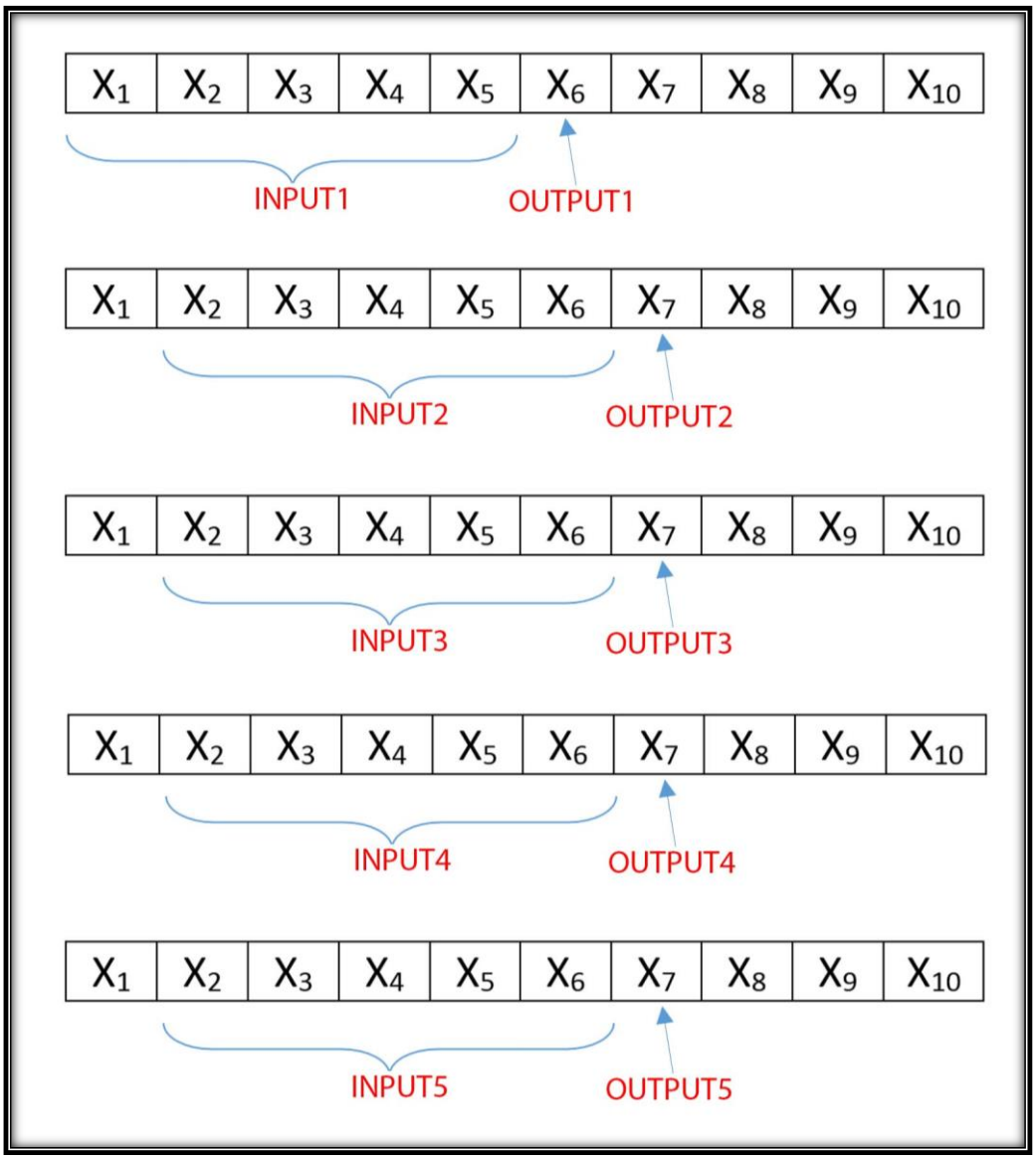
از روی ورودی های خام کافی است یک پنجره لغزان را بر روی ورودی خام حرکت دهید و در هر گام حرکت، داده های درون پنجره را به عنوان ورودی و اولین داده پس از پنجره را به عنوان خروجی در نظر بگیرید. با این کار ورودی ها و خروجی های پیش‌بینی‌گر MLP ساخته می‌شود. پنجره لغزان تا زمانی به حرکت خود ادامه می‌دهد که بتواند خروجی سیستم را مشخص نماید.

در پیش پردازش بالا بر روی ورودی های خام، ورودی های آموزش و تست شبکه عصبی ساخته می شود. بنابراین با توجه به مطالب بالا، پیش بینی گرMLP مورد استفاده در این پایان نامه، دارای ورودی با تعداد نرون به طول پنجره لغزان، یک لایه مخفی با تعداد نرون به طول پنجره لغزان و همچنین لایه خروجی با تنها یک نرون می‌باشد.

برای مثال فرض کنید ورودی خام، مجموعه $\{x_1, x_2, x_3, \ldots, x_{10}\}$ باشد و طول پنجره لغزان برابر ۴ باشد. در شکل۳- ۱ نمایش گرافیکی پنجره لغزان آورده شده است. نهایتا جدول۳- ۱ ورودی و خروجی های مربوط به آموزش شبکه عصبی را نمایش می دهد.

جدول۳- ۱:ورودی و خروجی های پیش بینی گر MLP برای ورودی خام $\{x_1, x_2, x_3, \ldots, x_{10}\}$ و طول پنجره لغزان ٤

خروجی	ورودی
$\mathbf{X_6}$	$\{X_1, X_2, X_3, X_4, X_5\}$
$\mathbf{X_7}$	$\{X_2, X_3, X_4, X_5, X_6\}$
$\mathbf{X_8}$	$\{X_3, X_4, X_5, X_6, X_7\}$
$\mathbf{X_9}$	$\{X_4, X_5, X_6, X_7, X_8\}$
$\mathbf{X_{10}}$	$\{X_5, X_6, X_7, X_8, X_9\}$



شکل۳- ۱:نمایش گرافیکی پنجره لغزان برای ورودی خام {x1, x2, x3,..., x10} و طول پنجره لغزان ۴

حال برای پیش بینی یک داده آینده، کافی است پنجره لغزان را در بخش پایانی ورودی های خام قرار داده و داده‌های داخل پنجره به عنوان ورودی به شبکه عصبی اعمال گردد. پاسخ شبکه عصبی به این ورودی به عنوان پیش بینی سیستم از داده همجوشی به مرحله بعدی فرستاده خواهد شد.

اندازه پنجره لغزان، وابسته به طول داده های خام می‌باشد. بهترین طول پنجره لغزان به کمک آزمایش بدست می‌آید. در این پایان نامه به منظور بدست آوردن طول پنجره، روش زیر پیشنهاد شده است. برای محاسبه طول پنجره لغزان کافی است اندازه گیری های یکی از منابع تا لحظه جاری را به عنوان ورودی داده خام انتخاب نمایید؛ سپس با تبدیل ورودی خام به ورودی شبکه عصبی با اندازه پنجره لغزان متفاوت و محاسبه خطای هر شبکه، بهترین طول پنجره را انتخاب نمایید. اگرچه در این روش یکی از منابع مورد بررسی قرار گرفته است اما امید می‌رود رفتار همه منابع به طور کلی مشابه باشد. ذکر این نکته مفید است که رفتار سیستم ما، شناسا نمی باشد.

نکته دیگری که در این بخش باید به آن توجه نمود این است که به طور مثال، در پیش بینی دمای هوا، رفتار سیستم در زمان های خیلی دور در پیش بینی تاثیر کمتری خواهد داشت. به همین منظور در این روش، پیشنهاد شده است تا داده های خام خیلی قدیمی حذف گردد. از آنجایی که یکی از شرایط سناریوی مطروح، طراحی برای سیستم‌های ناآشنا بوده است بنابراین نمی‌توان با مدل ریاضی، زمان تعویض رفتار را پیش بینی نمود.

۳-۲-۳- الگوریتم به روز رسانی پیش بینی

برای به روز رسانی پیش بینی، با داشتن m سنسور غیر قطعی، x_f به عنوان حاصل همجوشی در نظر گرفته می شود؛

$$x_f = \sum_{i=1}^{m} W_i \frac{1}{\alpha + |\hat{Z} - z_i|} \qquad (۱-۳)$$

در رابطه بالا، z_i مشاهده سنسور i ام و W_i وزن هر مشاهده می باشد. $\hat{Z}$ پیش بینی از مرحله قبل و α ضریب تاثیر این پیش بینی است. بردار وزن ها به شکل یک بردار سطری است.

$$W = [W_1 \quad W_2 \quad \dots \quad W_m] \qquad (۲-۳)$$

در رابطه بالا شرط زیر باید ارضا شود.

$$\sum_{i=1}^{m} W_i = 1 \qquad (۳-۳)$$

برای ارضای شرط بالا W_i به صورت زیر انتخاب شده است.

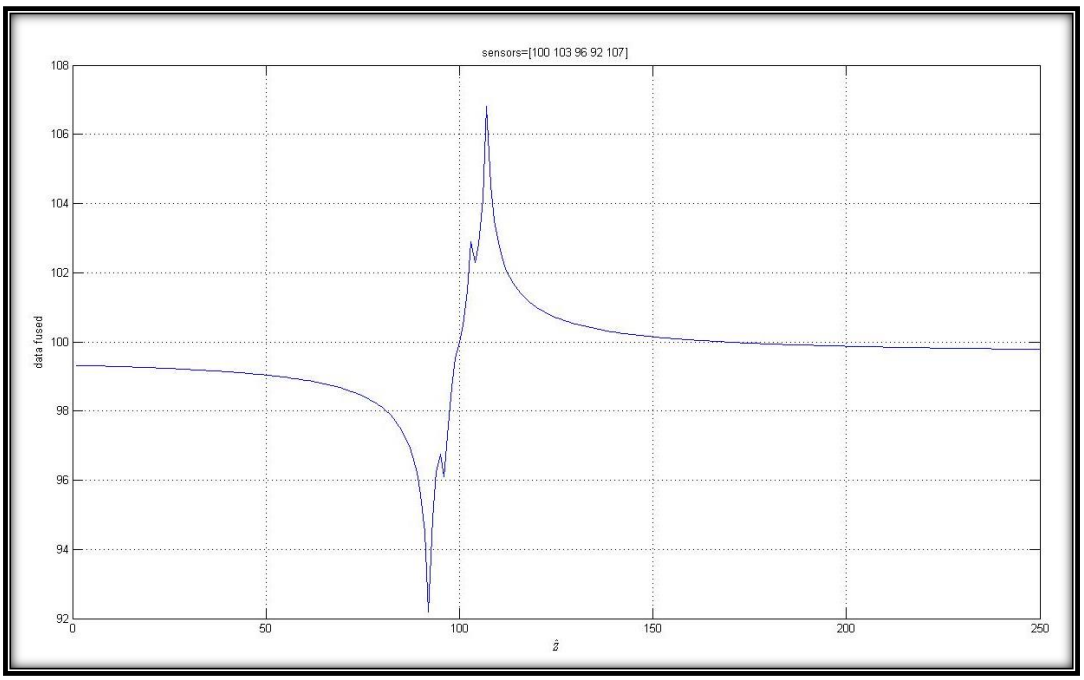
$$W_i = \frac{1}{\alpha + \left|\hat{Z} - z_i\right|} * \frac{1}{\sum_{j=1}^{m} \frac{1}{\alpha + \left|\hat{Z} - z_j\right|}} \qquad (۴-۳)$$

۳-۳- بررسی نقش پیش بینی گر

این بخش با چند مثال، به بررسی نقش پیش بینی در الگوریتم ارائه شده پرداخته است؛ در ابتدا فرض کنید ۵ منبع وجود دارد که بردار زیر نمایانگر مشاهدات آنها است.

Sensors= {100, 103, 96, 92, 107} (بردار ۳-۱)

شکل۳- ۲ ، نمودار همجوشی منابع بر اساس داده پیش بینی شده از مرحله قبل است. در این نمودار محور افقی، حاصل پیش بینی گر MLP است و محور عمودی، نتیجه همجوشی منابع با توجه به این پیش‌بینی می‌باشد. این نمودار نشان می دهد که اگر داده پیش بینی شده، خود دارای تناقض با سنسورها باشد در همجوشی داده ها کم تاثیر است؛ در حالیکه اگر این پیش بینی با خروجی های سنسور ها تطابق داشته باشد، تاثیر پیش‌بینی بیشتر خواهد بود.



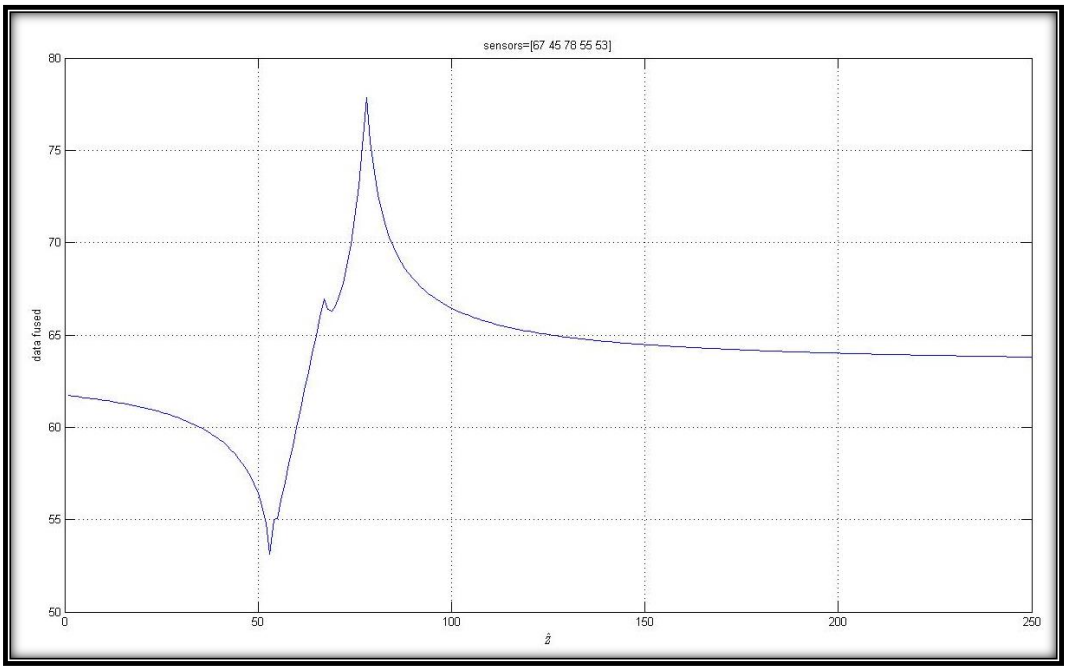
شکل۳- ۲: همجوشی منابع بر اساس پیش بینی MLP برای مشاهدات نزدیک به یکدیگر

در واقع اگر پیش بینی صحیح نباشد، الگوریتم خودش پیش بینی را حذف نموده و بدون توجه به آن تصمیم گیری می نماید. آزمایش بالا را با مشاهدات دیگری تکرار می کنیم، این مشاهدات

شباهت کمتری به یکدیگر داشته و بیشترین اختلاف بین آنها زیاد است. انتظار می‌رود نتیجه شبیه سازی مانند قبل باشد. شکل۳- ۳، این انتظار را برآورده ساخته است. تفاوتی که در بین این دو نمودار وجود دارد، این است که محدوده ایفای نقش زیاد پیش‌بینی‌گر در منابع دور از هم، گسترده تر از منابع نزدیک به همدیگر است. این ویژگی جالبی است زیرا هنگامی که مشاهدات منابع نزدیک به همدیگر هستند نیاز زیادی به پیش بینی گر نیست. در پردازش روی کلاستر ها گفته شد اگر بیش از نیمی از مشاهدات منابع همدیگر را تایید کردند بدون توجه به پیش بینی گر همجوشی انجام می شود. در اینجا نیز، نقش پیش بینی گر را کنترل شده است که چه زمانی پررنگ تر و چه زمانی کم رنگ تر در همجوشی ظاهر شود. بخش پردازش روی کلاستر ها، یک حد آستانه[1] برای این نقش می‌باشد.

Sensors= {67, 45, 78, 55, 53} (بردار ۳-۲)

[1] threshold



شکل۳- ۳:همجوشی منابع بر اساس پیش بینی MLP برای مشاهدات دور از یکدیگر

مسئله دیگری که باید در این بخش به آن پرداخته شود، نقش مقدار پیش بینی بر روی وزن های الگوریتم به روز رسانی است. در الگوریتم مربوط به بخش به‌روزرسانی مقدار پیش‌بینی، هرچه مشاهده یا اندازه‌گیری یک منبع به مقدار پیش‌بینی شده نزدیک‌تر باشد، وزن تاثیر آن منبع نیز بیشتر می‌شود. به عبارت دیگر منابعی که مشاهده‌ی نزدیک‌تری به مقدار پیش‌بینی دارند با ارزش‌ترند و احتمالاً مقدار عدم قطعیت و تناقض آن‌ها کمتر است. این ادعا در ادامه اثبات شده است.

فرض کنید مقدار مشاهده شده توسط منبع **j** به مقدار پیش‌بینی شده نزدیکتر از مقدار مشاهده شده توسط منبع **i** به مقدار پیش‌بینی شده باشد یا به عبارت دیگر:

$$|z_i - \hat{Z}| > |z_j - \hat{Z}| \qquad (۵-۳)$$

بنابراین با فرض مثبت بودن ضریب تاثیر، می توان نتیجه گرفت که:

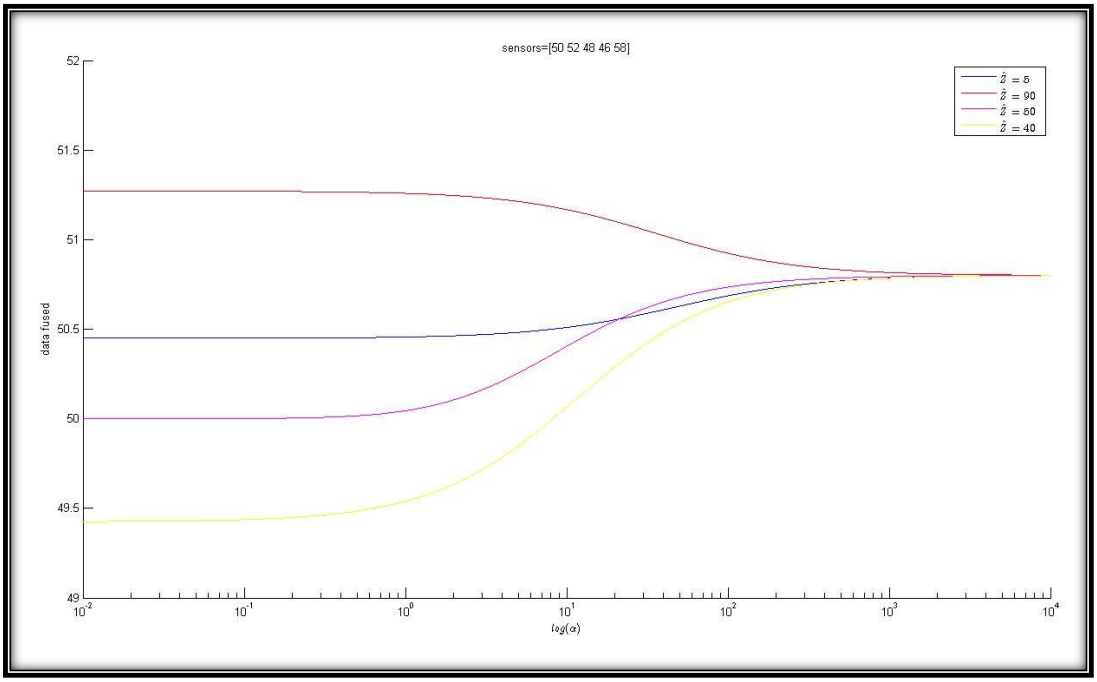
$$\frac{1}{\alpha + |\hat{Z} - z_i|} < \frac{1}{\alpha + |\hat{Z} - z_j|} \qquad (۶-۳)$$

بنابراین ادعای بالا ثابت شده است و داریم:

$$w_i < w_j \qquad (۷-۳)$$

۳-۴- بررسی نقش ضریب تاثیر

در این بخش از پایان نامه با در نظر گرفتن مشاهدات ۵ منبع، نتیجه همجوشی با ضرایب تاثیر متفاوت رسم شده است. شکل۳- ۴ نشان می دهد نتیجه همجوشی به ضریب تاثیر وابسته است. در شکل۳- ۴، محور افقی، به شکل لگاریتمی و مقدار ضریب آلفا را نشان می دهد. محور عمودی نتیجه همجوشی متناظر با ضریب تاثیر است.



شکل۳- ٤: نقش ضریب تاثیر بر روی نتیجه همجوشی

شکل۳- ۴، نشان می دهد که با افزایش ضریب تاثیر، همجوشی منابع به میانگین مشاهدات نزدیک می‌شود و هر چه مقدار ضریب تاثیر کمتر شود همجوشی داده ها با نگاه بیشتری به عدم قطعیت محاسبه می شود.

۳–۵– اعمال الگوریتم پیشنهاد شده بر روی داده های هواشناسی

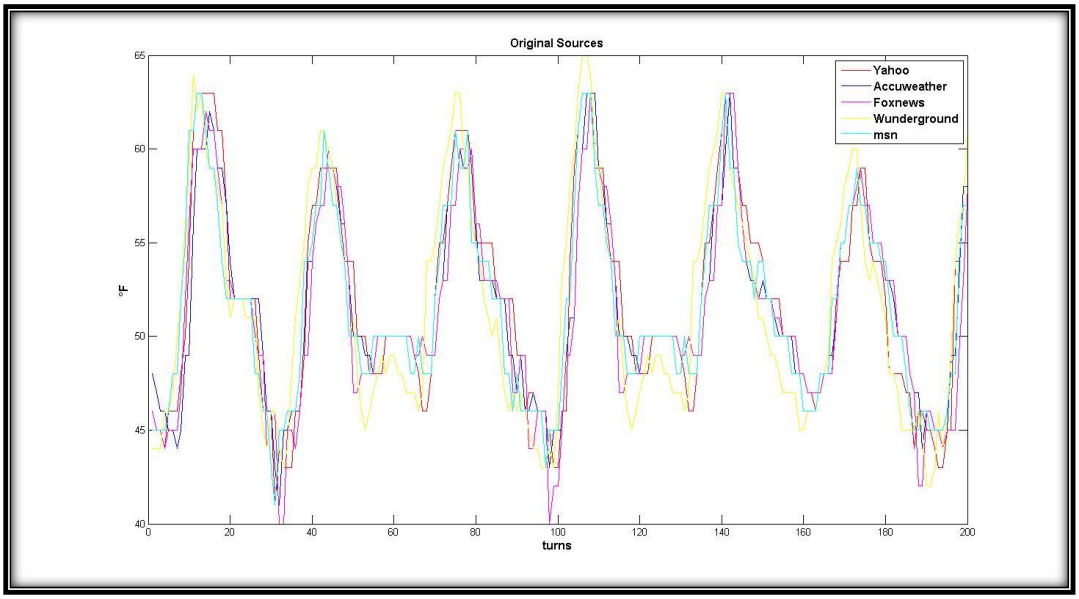
در این بخش، الگوریتم پیشنهاد شده در پایان نامه، بر روی یک داده واقعی اعمال شده است. به همین منظور از منبع [26] مجموعه داده[1]ای گرفته شده است. در این مجموعه داده، دماهای اعلام شده توسط ۱۸ وبسایت برای ۳۰ شهر بزرگ آمریکا در مارس ۲۰۱۰ جمع آوری شده است. این دماها هر ۴۵ دقیقه یکبار از این سایت‌ها استخراج شده است. هیچکدام از این وبسایت ها در

[1] dataset

این مناطق جغرافیایی سنسور فیزیکی دما نداشته و فقط پیش بینی خود را از راه دور اعلام می کنند. پیش بینی این شرکت ها در بسیاری از موارد همدیگر را تایید می کنند اما در مواردی این پیش بینی ها موئد همدیگر نیستند. هدف این پایان نامه ارائه الگوریتمی کارآ برای شناسایی و رسیدگی به عدم قطعیت و تناقضات در این نمونه داده ها می باشد.

۳-۵-۱- همجوشی داده ها با ۵ منبع

از مجموعه داده [26] ، داده های مربوط به ۵ شرکت برای شهر San Jose از کالیفرنیا انتخاب شده است. این ۵ شرکت عبارتند از: foxnews ، yahoo ، wunderground ، accuweather و msn. با توجه به هدف الگوریتم ارائه شده در این پایان نامه، آن الگوریتم به این داده‌ها اعمال شده است. در شکل۳- ۵، دماهای اندازه‌گیری شده توسط این پنج شرکت رسم شده است.



شکل۳- ۵: دمای گزارش شده برای شهر San Jose

با توجه به نمودار بالا، مشخص است که در همه زمان ها پیش بینی ها همدیگر را تایید نمی کنند، اگر به داده‌های نوبت ۶۰ دقت شود، مشخص می شود که به جز شرکت wunderground ، همه شرکت ها هم نظر هستند و این شرکت به اشتباه پیش بینی کرده است.

S[60]= {50, 50, 50, 49, 50} (بردار ۳-۳)

ذکر این نکته در این بخش خالی از لطف نمی‌باشد که تناقض یا عدم قطعیت همواره معنای ثابتی ندارند؛ به عبارت دیگر یک درجه فارنهایت اختلاف در بالا، یک تناقض بوده است در حالی که در مورد زیر یک درجه اختلاف به معنای داده متناقض نخواهد بود. در اینجا به وضوح مشخص است دمای ۵۵ و ۵۲ درجه فارنهایت دارای مقداری خطا می باشد.

Sensors= {55, 53, 53, 50, 52} (بردار ۳-۴)

قبل از همجوشی منابع بالا، در ابتدا با کمک رابطه زیر، میانگین مربعات فاصله(MSD) منابع سنجیده شده است. MSD بین دو سیگنال s و v از رابطه زیر بدست می‌آید.

$$MSD(s,v) = \frac{1}{N}\sum_{i=1}^{N}(s_i - v_i)^2 \qquad (۸-۳)$$

با توجه به شکل۳- ۵، دماهای گزارش شده توسط شرکت wunderground از بقیه شرکت ها متفاوت‌تر است. جدول۳- ۲، نیز نشان می‌دهد منبع wunderground اختلاف بیشتری با بقیه منابع دارد. توقع داریم که نتیجه همجوشی به بقیه منابع نزدیک‌تر باشد.

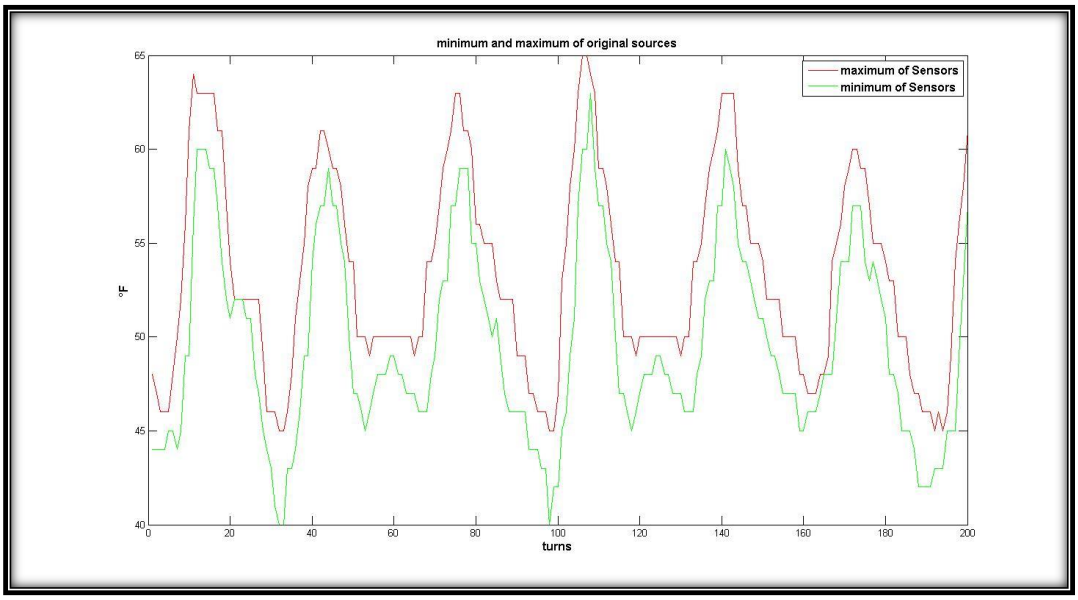
جدول۳- ۲: میانگین مربعات خطا بین منابع برای شهر San Jose

	yahoo	*accuweather*	*foxnews*	*wunderground*	*msn*
yahoo	0	4.045	5.535	8.435	4.49
accuweather	4.045	0	2.6	10.87	4.885
foxnews	5.535	2.6	0	11.83	5.075
wunderground	8.435	10.87	11.83	0	5.475
msn	4.49	4.885	5.075	5.475	0

در شکل۳- ۶، حداکثر و حداقل دمای گزارش شده برای شهر San Jose نشان داده شده است. این نمودار نشان می دهد گاهی همه منابع، دقیقا یک پیش بینی را ارائه می کنند(در نوبت ۲۱) و گاهی منابع، با اختلاف زیادی مانند ۹ درجه فارنهایت، پیش بینی های خود را ارائه می نمایند(در نوبت ۳۹).

$$S[21]= \{52, 52, 52, 52, 52\} \quad \text{(بردار ۳-۵)}$$

$$S[39]= \{55, 54, 49, 58, 54\} \quad \text{(بردار ۳-۶)}$$

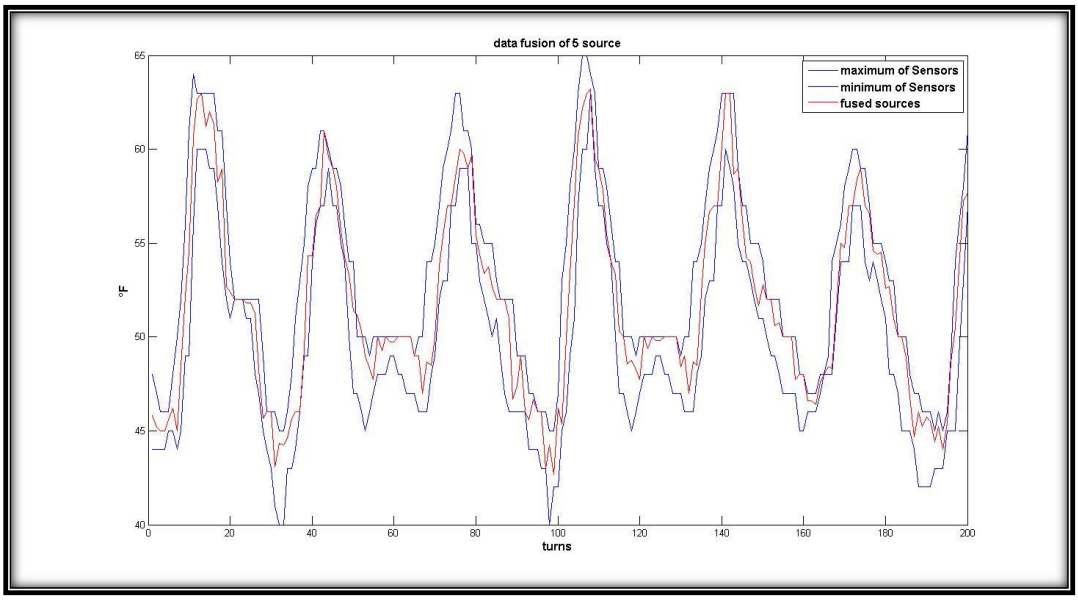


شکل۳- ۶:حداکثر و حداقل دمای گزارش شده برای شهر San Jose

با الگوریتم ارائه شده در این پایان نامه، همجوشی داده ها محاسبه شده است. نتیجه همجوشی در شکل۳- ۷ آورده شده است. در بالا، انتظار می‌رفت که نتیجه همجوشی منابع از wunderground دورتر بوده و به بقیه منابع نزدیک‌تر باشد. جدول۳- ۳ نشان می‌دهد که الگوریتم ارائه شده این انتظار را برآورده ساخته است.

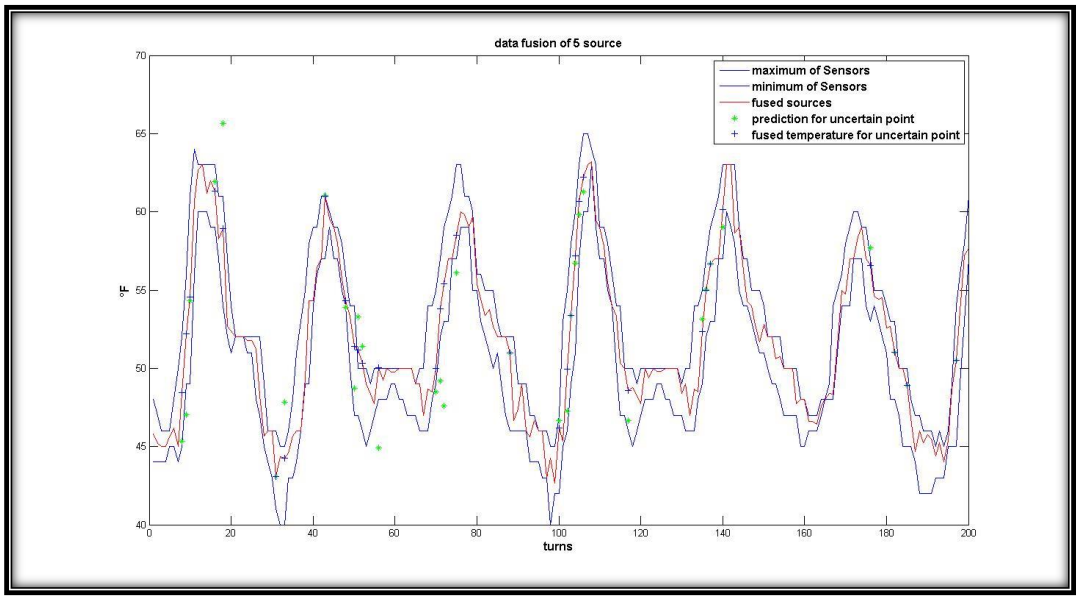
جدول۳- ۳: میانگین مربعات خطا بین همجوشی منابع و تک تک منابع برای شهر San Jose

	yahoo	*accuweather*	*foxnews*	*wunderground*	*msn*
MSD	2.463	1.856	3.133	7.550	2.644



شکل۳- ۷:همجوشی منابع برای شهر San Jose

همانطور که قبلا به آن اشاره شده است منابع در بسیاری از نقاط مؤید یکدیگر هستند. این تایید با توجه به الگوریتم مربوط به بخشی است که بیشتر از نیمی از داده ها در یک کلاستر قرار بگیرند. بنابراین این نقاط نیز دارای سطحی از تناقض یا عدم قطعیت می باشند. در این پایان نامه، سایر نقاط که این شرایط را ندارند نقاط تناقض و عدم قطعیت شدید نامگذاری شده است. در شکل۳- ۸ این نقاط نشان داده شده است. در شکل۳- ۸، ستاره های سبز رنگ با شبکه عصبی پیش بینی شده و نقاط با علامت جمع آبی، نتیجه همجوشی پیش بینی با مقادیر سنسور ها می باشد.



شکل۳- ۸:همجوشی منابع برای شهر San Jose با جزییات نقاط تناقض و عدم قطعیت شدید

مشخصات نوبت ۱۸ در جدول۳- ۴ آمده است، پیش بینی شبکه عصبی، دارای مقداری خطا می‌باشد و به نظر می‌رسد باید خود این پیش بینی کم تاثیر شود. قبلا گفته شد اگر پیش بینی شبکه عصبی دارای خطا باشد نتیجه همجوشی به سمت میانگین گزارشات منابع رفته و تاثیرات داده پیش بینی، کمتر خواهد شد. اگر به گزارشات منابع توجه نماییم نتیجه همجوشی به همجوشی ذهن انسان نزدیک است. همچنین جدول۳- ۵ نشان می‌دهد اگر داده پیش بینی شده با شبکه عصبی به منابع نزدیک باشد، نتیجه همجوشی به میانگین وزن دار منابع نزدیک می‌شود.

جدول۳- ٤: مشخصات نوبت ۱۸ برای شهر San Jose

نتیجه همجوشی	**۵۸,۹۵**
داده پیش بینی شده با شبکه MLP	٦٥,٦٤
گزارشات منابع	{۶۱،۵۹،۵۷،۵۷،۵۴}
میانگین گزارشات منابع	۵۷,٦

جدول۳- ۵: مشخصات نوبت ۱۰۰ برای شهر San Jose

نتیجه همجوشی	**٤٦,٢٠**
داده پیش بینی شده با شبکه MLP	٤٦,٦٦
گزارشات منابع	{۴۳،۴۵،۴۲،۴۷،۴۵}
میانگین گزارشات منابع	٤٤,٤٠

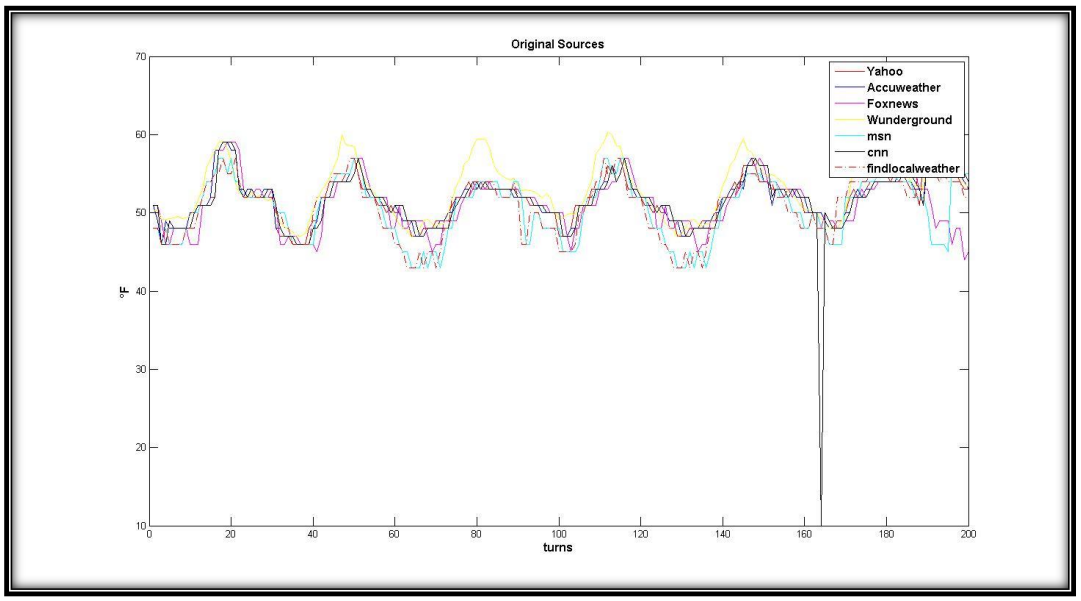
۳-۵-۲- همجوشی داده ها با ۷ منبع

در این بخش، از مجموعه داده [26] آزمایش بالا برای شهر San Francisco با ۷ منبع تکرار شده است. علاوه بر منابع قبلی، گزارشات دو منبع cnn و findlocalweather برای همجوشی اضافه شده است. در بخش بعدی به کمک نتیجه این بخش، قیاسی بین نتایج همجوشی داده ها بین ۵ و ۷ منبع آورده شده است. با توجه به شکل۳- ۹ که دماهای گزارش شده از ۷ منبع در آن نشان داده شده است، شاید نتوان به دقت توسط مغز پیش‌بینی نمود که نتیجه همجوشی به کدام منابع نزدیک‌تر است. جدول۳- ۶ نشان می‌دهد منابع msn و cnn از بقیه منابع فاصله بیشتری دارند؛ بنابراین توقع داریم نتیجه همجوشی از msn و cnn دور بوده و به بقیه منابع نزدیک باشد. همجوشی داده ها با ۷ منبع برای شهر San Francisco در شکل۳- ۱۰ آورده شده است. با محاسبه مقادیر میانگین مربعات فاصله در جدول۳- ۷ مشاهده می‌شود همجوشی ۷ منبع از منابع msn و cnn دور بوده و به بقیه منابع نزدیک تر است.

جدول۳- ٦: میانگین مربعات فاصله بین منابع برای شهر San Francisco

	yahoo	*accuweather*	*foxnews*	*wunderground*	*msn*	*cnn*	*Findlocalweather*
yahoo	0.00	1.03	5.76	5.24	7.81	9.38	4.95
accuweather	1.03	0.00	5.67	5.32	7.97	8.82	5.61
foxnews	5.76	5.67	0.00	11.99	8.56	12.39	9.20
wunderground	5.24	5.32	11.99	0.00	12.48	14.04	8.73

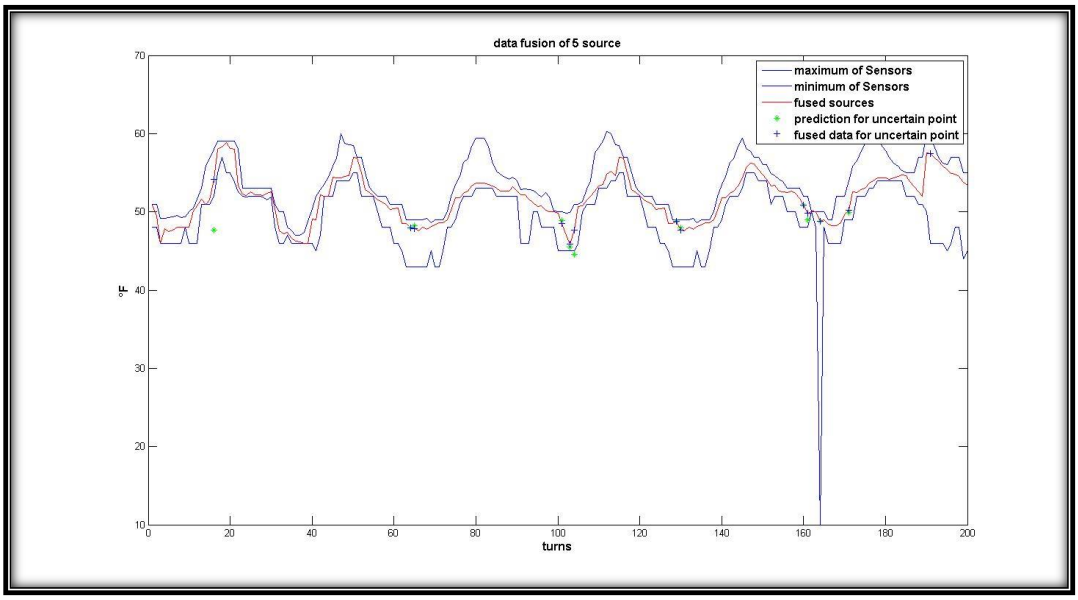
msn	7.81	7.97	8.56	12.48	0.00	15.97	4.77
cnn	9.38	8.82	12.39	14.04	15.97	0.00	13.88
findlocalweather	4.95	5.61	9.20	8.73	4.77	13.88	0.00



شکل۳- ۹:دمای گزارش شده برای شهر San Francisco

جدول۳- ۷: میانگین مربعات فاصله بین همجوشی منابع و تک تک منابع برای شهر San Francisco

	yahoo	*accuweather*	*foxnews*	*wunderground*	*msn*	*cnn*	*findlocalweather*
MSD	0.712	0.401	4.857	5.106	7.379	8.057	5.158



شکل۳- ۱۰:همجوشی منابع برای شهر San Francisco با جزییات نقاط تناقض و عدم قطعیت شدید

مشخصات نوبت ۱۶۴ در جدول۳- ۸ آورده شده است. از آنجایی که پیش بینی منبع cnn دارای تناقض با سایر منابع است، این منبع از همجوشی حذف شده و با بقیه منابع و میانگین وزن دار آن‌ها نتایج همجوشی محاسبه شده است.

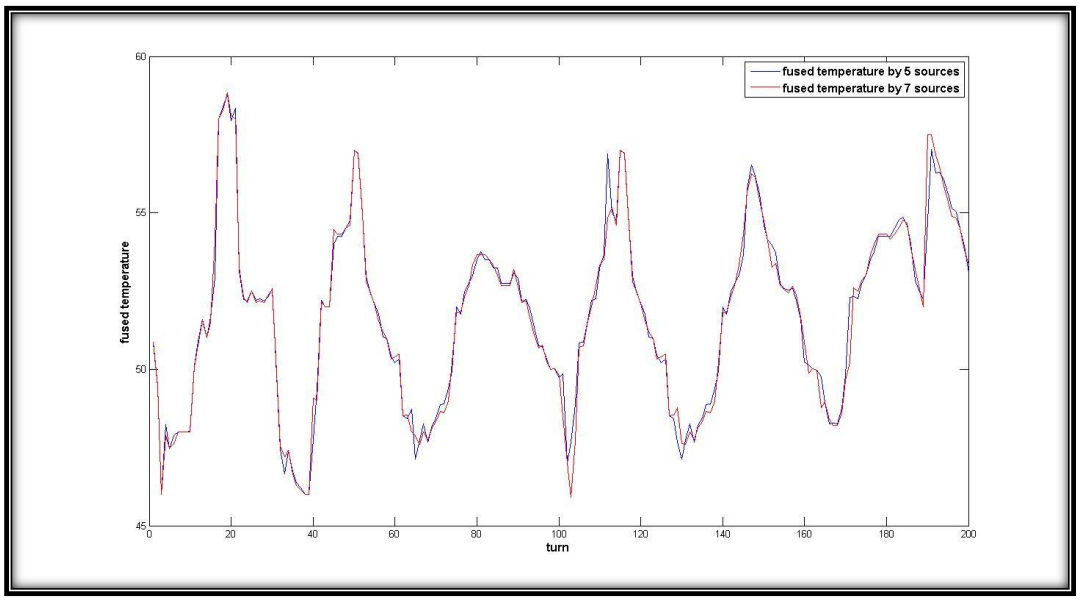
جدول۳- ۸: مشخصات نوبت ۱۶۴ برای شهر San Francisco

نتیجه همجوشی	**۴۸٫۷۸**
داده پیش بینی شده با شبکه MLP	٤٨,٧٨
گزارشات منابع	{۵۰،۵۰،۴۸،۴۹/۳،۴۸،۱۰،۴۸}
میانگین گزارشات منابع	٤٣,٣٣

۳-۵-۳- مقایسه همجوشی ۵ و ۷ منبع

در این بخش از پایان نامه، هدف قیاس نتیجه همجوشی با ۵ و ۷ منبع می‌باشد. نتایج همجوشی با ۷ منبع برای شهر San Francisco در بخش قبلی آورده شده است. با حذف منابع cnn و findlocalweather از آزمایش قبلی مربوط به شهر San Francisco ، آزمایش با ۵ منبع دوباره تکرار شده است.

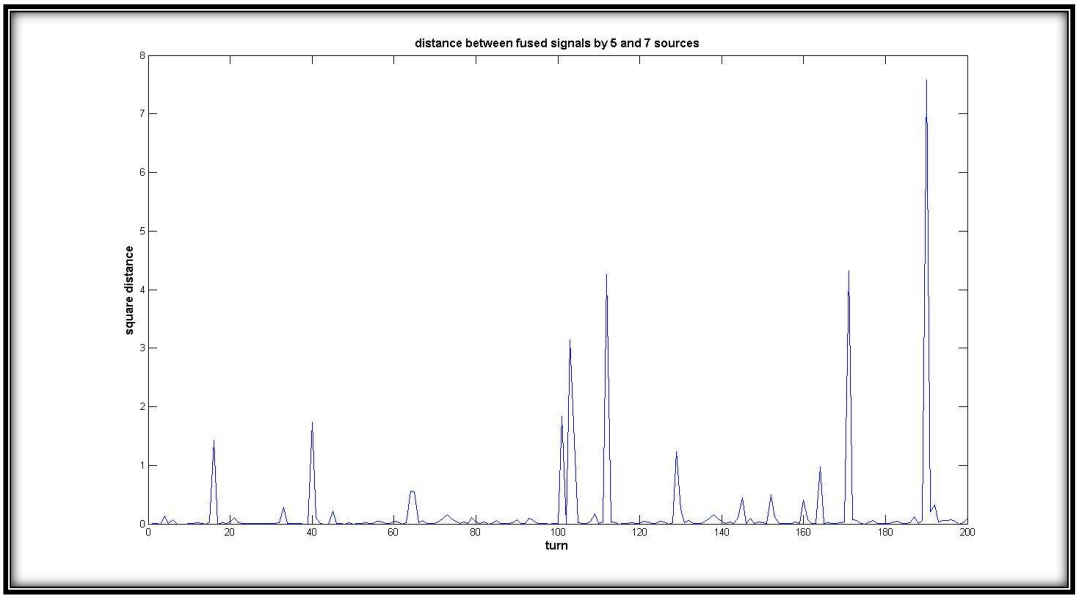
در شکل۳- ۱۱، همجوشی منابع با ۷ و ۵ منبع رسم شده اند. همانطور که مشهود است این دو همجوشی در بسیاری از نقاط هم مقدار یا نزدیک به همدیگر است. برای درک بهتر از میزان شباهت این دو، میزان مربعات فاصله آن‌ها نیز در شکل۳- ۱۲ رسم شده است.



شکل۳- ۱۱: همجوشی دما با ۵ و ۷ منبع برای شهر San Francisco

$$SD\{s,v\}(i,j) = (s(i) - v(j))^2 \qquad (۹-۳)$$

که در رابطه بالا، مربعات فاصله بین دو نقطه S_i و V_j در دو سیگنال S و V محاسبه شده است. با کمک همین رابطه، مربعات فاصله بین دو سیگنال همجوشی شده با ۵ و ۷ منبع محاسبه شده است.



شکل۳- ۱۲:میزان مربعات فاصله بین همجوشی ها با ۵ و ۷ منبع

در شکل۳- ۱۲ میزان مربعات فاصله بین دو سیگنال همجوشی شده با ۵ و ۷ منبع آورده شده است. در همین آزمایش میانگین مربعات فاصله محاسبه و مقدار آن برابر ۰,۱۸۰ می‌باشد. این مقدار نشان می‌دهد که این دو سیگنال به یکدیگر شبیه هستند. با توجه به شکل مشخص می‌شود بیشتر این میانگین بخاطر چند نقطه خاص در دو سیگنال به وجود آمده است. جدول۳- ۹ مشخصات نوبت ۱۹۰ که دارای فاصله زیادی است را نشان می‌دهد. در همجوشی با ۵ منبع، گزارشات منابع کاملا متفاوت هستند و همجوشی با پیش‌بینی شبکه عصبی و به روز رسانی آن‌ها انجام شده است. این نتیجه همجوشی به نظر منطقی است. در همجوشی با ۷ منبع، دو منبع اضافه شده با تایید دو گزارش از منابع قبلی، منابع متناقض را مشخص می سازد و با حذف این

منابع نتیجه همجوشی محاسبه شده است. تامل در جدول۳- ۹ نشان می‌دهد که علت فاصله زیاد بین دو همجوشی با ۵ و ۷ منبع در نقطه ۱۹۰، عوض شدن ماهیت گزارشات در همجوشی با ۷ منبع می‌باشد.

جدول۳- ۹: مشخصات نوبت ۱۹۰ برای شهر San Francisco

نتیجه همجوشی با ۵ منبع	**۵۴/۷۵**
نتیجه همجوشی با ۷ منبع	**۵۷,۵۰**
گزارشات منابع برای ۵ منبع	**{۵۸،۵۸،۵۳،۶۰،۵۰}**
گزارشات منابع برای ۷ منبع	**{۵۸،۵۸،۵۳،۶۰،۵۰،۵۷،۵۷}**
فاصله بین دو همجوشی	**۲,۷۵**
مربع فاصله بین دو همجوشی	**۷,۵۶**

جدول۳- ۱۰ نیز، نشان دهنده مشخصات نقطه دیگری با فاصله زیاد بین دو همجوشی با ۵ و ۷ منبع می‌باشد. در پیش‌بینی با ۵ منبع، سه منبع با پیش‌بینی‌های ۵۲، ۵۲,۹ و ۵۲ بیان می‌دارند که نتیجه همجوشی با ۵ منبع، نزدیک به این داده ها می باشد. طبق الگوریتم ارائه شده در این پایان نامه، پس از خوشه بندی، اگر بیش از نیمی از منابع در خوشه قرار بگیرند نتیجه همجوشی برابر میانگین این خوشه خواهد بود. به هر حال، نتیجه همجوشی ۵ منبع کاملا به ذهن انسان نزدیک است.

جدول۳- ۱۰: مشخصات نوبت ۱۷۱ برای شهر San Francisco

نتیجه همجوشی با ۵ منبع	**۵۲/۳۰**
نتیجه همجوشی با ۷ منبع	**۵۰,۲۲**
گزارشات منابع برای ۵ منبع	**{۵۲،۵۰،۴۹،۵۲/۹،۵۲}**
گزارشات منابع برای ۷ منبع	**{۵۲،۵۰،۴۹،۵۲/۹،۵۲،۵۰،۵۴}**

فاصله بین دو همجوشی	۲,۰۷
مربع فاصله بین دو همجوشی	٤,۲۸

برای همجوشی با ۷ منبع، ابتدا خوشه بندی انجام می شود. نتایج خوشه بندی در جدول۳- ۱۱، منعکس شده است. خوشه بندی ها در سه دسته انجام شده اند. دو دسته دارای ۳ عضو و یک دسته دارای ۱ عضو می‌باشد. همانطور که در الگوریتم ارائه شده مطرح شد، خوشه با یک عضو از همجوشی حذف خواهد شد. این کار به ذهن انسان هم نزدیک است. اگرچه در این نقطه نیز، فاصله بین دو همجوشی زیاد است اما بخاطر تغییر ماهیت گزارشات می باشد نه ضعف در الگوریتم. در واقع با افزوده شدن دو منبع آخر، احتمال نزدیکی دما به دسته دماهای حدود ۵۰ درجه فارنهایت افزایش می یابد.

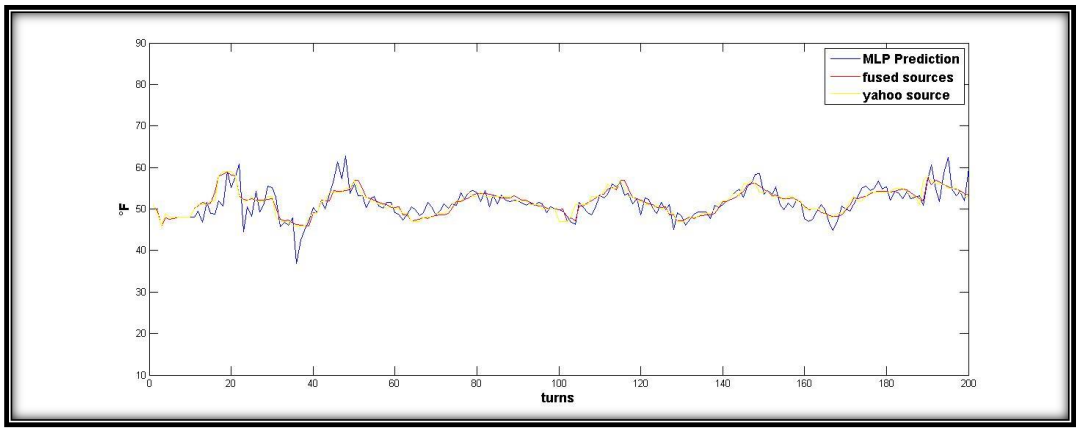
جدول۳- ۱۱: خوشه بندی گزارشات ۷ منبع برای نوبت ۱۷۱ از شهر San Francisco

گزارشات منابع برای ۷ منبع	**{۵۲،۵۰،۴۹،۵۲/۹،۵۲،۵۰،۵۴}**
خوشه اول	{۵۲،۵۲/۹،۵۲}
خوشه دوم	{۵۰،۴۹،۵۰}
خوشه سوم	{۵۴}

۳–۵–۴– مقایسه همجوشی منابع با پیش بینی تک منبعی

در این بخش، با روش شبکه MLP و یکی از منابع، پیش بینی دما انجام شده است. سپس همان دما با کمک همجوشی محاسبه گردیده است. نتایج نشان می دهد که روش ارائه شده در این پایان نامه، از روش پیش بینی با شبکه MLP تواناتر است.

برای یک هفته از سال، دمای هوا با شبکه MLP پیش بینی شده است. برای آموزش شبکه در هر مرحله از گزارشات دما و پنجره لغزان استفاده می شود. دمای هوا هر ۴۵ دقیقه یکبار گزارش گیری شده است. نتایج آزمایش برای شهر San Francisco در شکل۳- ۱۳ آمده است. شبکه MLP در پیش بینی لحظات ابتدایی بسیار ناتوان است زیرا این شبکه برای آموزش نیاز به داده های بیشتری نسبت به روش ارائه شده دارد. همانطور که در شکل۳- ۱۳ مشهود است نتیجه همجوشی نزدیکی زیادی به ذهن انسان دارد. این مقایسه برای نقاط مختلف انجام شده است که در این پایان نامه از آوردن آن خودداری شده است. نکته قابل توجه دیگر، سرعت بالای الگوریتم ارائه شده می‌باشد. الگوریتم ارائه شده با حذف مرحله پیش بینی در بعضی نقاط سرعت اجرا را بهبود بخشیده است.



شکل۳- ۱۳: مقایسه نتیجه همجوشی با شبکه MLP

۳-۶- طراحی سناریو برای شناسایی آتش سوزی

در بخش قبلی، الگوریتم ارائه شده در این پایان نامه بر روی داده های هواشناسی مورد استفاده قرار گرفت و نتایج آن مورد بررسی و مطالعه قرار گرفت. در این بخش، به دنبال نگاه دیگری به الگوریتم ارائه شده در این پایان نامه هستیم. سناریوی زیر را در نظر بگیرید.

یکی از راه های شناسایی آتش سوزی در جنگل ها، پراکنده سازی منابع اندازه گیری و سنسور ها در محیط جنگل می‌باشد. به مجموعه ی این منابع، WSN نیز گفته می شود. این منابع دارای انرژی محدود هستند و ارسال دائم اطلاعات توسط آن ها باعث کاهش عمرشان می شود. فرض کنید برای شناسایی یک جنگل شبکه حسگر های بیسیم را در محیط ایجاد نموده اید. یک سرور برای همجوشی مشاهدات حسگر ها لازم است و طبیعتا همه سنسور ها می خواهند مشاهدات خود را به سرور انتقال دهند. انتقال همه مشاهدات باعث مصرف بالای انرژی می شود. سرور باید مدیریت نماید که چه زمانی نیاز به دریافت مشاهدات منابع دارد.

برای حل این مسئله، می توان از بخشی از الگوریتم ارائه شده در این پایان نامه استفاده نمود. اگرچه تاکید روش ارائه شده، برای این گونه مصارف نمی باشد، اما می توان آنرا به این مدل سناریو ها نیز اعمال نمود. برای شناسایی آتش سوزی، در ابتدا سرور به تعداد مشخصی، مشاهدات را دریافت می‌نماید. در صورتی که سرور با روش خوشه‌بندی مطرح شده یقینا به جواب این پرسش رسید که آیا آتش سوزی رخ داده است یا نه، فعالیت سرور به پایان می‌رسد. در غیر این صورت، سرور تقاضا می دهد که منابع دیگر نیز مشاهدات خود را ارسال نمایند. با این کار تعداد ارسال ها کاهش یافته و باعث ذخیره انرژی و افزایش عمر سنسور ها می شود.

فصل چهارم – نتیجه گیری و پیشنهادات برای کارهای آینده

در این پایان نامه، یک الگوریتم برای همجوشی داده ها در شرایط عدم قطعیت ارائه شده است. در این پایان نامه، فرض شده است که معادلات ریاضی رفتاری سیستم در اختیار ما قرار ندارد. الگوریتم ارائه شده، شامل تکنیک خوشه‌بندی، شبکه عصبی و قانون جدید به روز رسانی می‌باشد. الگوریتم پیشنهاد شده بر روی داده های هواشناسی اعمال شده است. آزمایشات انجام شده، نشان می دهد که این الگوریتم، توانایی شناسایی و برخورد با عدم قطعیت و تناقضات را دارد. علاوه بر داده های هواشناسی، این الگوریتم توانایی پیاده سازی بر روی هر مشاهدات یا اندازه گیری هایی با یکا های یکسان را دارد. به عبارت دیگر، الگوریتم پیشنهاد شده در این پایان نامه، فقط توانایی اعمال بر روی منابع همجنس را دارد. بنابراین به طور مثال، این الگوریتم توانایی همجوشی برای گزارشات مربوط به دمای هوا و سرعت باد را ندارد. الگوریتم پیشنهاد شده، می تواند برای نرم افزار های موبایل، نرم افزار های خانگی و وبسایت ها مورد استفاده قرار بگیرد.

با توجه به سناریوی دیگری که برای شناسایی آتش سوزی طراحی شده است، بخش ابتدایی الگوریتم می‌تواند برای همجوشی با حداقل ارسال مشاهدات مورد استفاده قرار گیرد. این ایده فقط با طرح اولیه آن در این پایان نامه مورد توجه قرار گرفته و می تواند به عنوان کارهای آتی، با شبیه سازی های مورد امتحان قرار بگیرد.

برای کارهای آتی، پیشنهاد می شود که الگوریتم پیشنهاد شده برای مشاهدات یا اندازه گیری های غیر همجنس گسترش یابد. اینکار می تواند با طراحی یک تابع تبدیل از مشاهدات و اندازه گیری های غیر همجنس به داده های همجنس انجام شود.

منابع

[1] B. Khaleghi, A. Khamis, F. O. Karray and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion,* vol. 12, pp. 28-44, 2013.

[2] L. A. Klein, Sensor and Data Fusion; A Tool for Information Assessment and Decision Making, Washington USA: The International Society for Optical Engineering, Third Printing 2010.

[3] W. A. Abdulhafiz and A. Khamis, "Handling Data Uncertainty and Inconsistency Using Multisensor Data Fusion," *Advances in Artificial Intelligence,* 2013.

[4] "LM35 Precision Centigrade Temperature Sensors," Texas Instruments, 2013.

[5] J. Willey, Uncertainty Theories and Multisensor Data Fusion, Great Britain: CPI Group, 2014.

[6] E. F. Nakamura, A. A. F. Loureiro and A. C. Frery, "Information fusion for wireless sensor networks: Methods, models, and classifications," *ACM Computing Surveys (CSUR),* vol. 39, no. 3, 2007.

[7] N. A. Carlson, "Federated square root filter for decentralized parallel processors," *Aerospace and Electronic Systems, IEEE Transactions,* vol. 26, no. 3, pp. 517-525, 1990.

[8] M. Kazerooni, F. Shabanin, M. Vaziri and S. Vadhva, "Federated ensemble Kalman filter in no reset mode design," in *International Conference on Information Reuse and Integration*, San Francisco, USA, 2013.

[9] J. Hou, Z. Jing, T. Gao and Y. Yang, "Adaptive phased-array tracking in the standoff jammer using H infinity filter," in *Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on*, Beijing, 2013.

[10] W. Naeem, R. Sutton and T. Xu, "An integrated multi-sensor data fusion algorithm and autopilot implementation in an uninhabited surface craft," *Ocean Engineering,* vol. 39, pp. 43-52, 2012.

[11] D. Fincher and D. Max, "Multisensor data fusion using neural networks," in *IEEE international Conference on Systems, Man and Cybernetics*, Los Angeles, USA, 1990.

[12] R. Ji-Xu, S. Jia-Cheng, H. Ji-Liu and F. Xiao-Yu, "Multi-sensor Fuzzy Stochastic Fusion Based on Genetic Algorithms," in *Information Acquisition, 2006 IEEE International Conference on*, Shandong, 2006.

[13] K. C. Chou, A. S. Willsky and A. Benvenste, "Multiscale recursive estimation, data fusion, and regularization," *Automatic Control, IEEE Transactions,* vol. 39, no. 3, pp. 464-478, 1994.

[14] D. Zhang, W. Xiaolin, P. Quan and Z. Hongcai, "Multiresolution modeling and estimation of multisensor data," *Signal Processing, IEEE Transactions,* vol. 52, no. 11, pp. 3170-3182, 2004.

[15] M. Kumar, D. P. Garg and R. A. Zachery, "A generalized approach for inconsistency detection in data fusion from multiple sensors," in *American Control Conference*, Minneapolis, MN, USA, 2006.

[16] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976.

[17] T. Khan, P. Ramuhalli and S. C. Dass, "Particle-Filter-Based Multisensor Fusion for Solving Low-Frequency Electromagnetic NDE Inverse Problems," *Instrumentation and Measurement, IEEE Transactions,* vol. 60, no. 6, pp. 2142-2153, 2011.

[18] M. Toufik and M. Nohad, "Partitions of a set satisfying certain set of conditions," *Discrete Mathematics,* vol. 309, no. 13, pp. 4481-4488, 2009.

[19] B. S. Everitt, S. Landau and M. Leese, Cluster Analysis, United Kingdom: Wiley, 2009.

[20] D. Lam and D. C. Wunsch, "Clustering," in *Signal Processing Theory and Machine Learning*, USA, 2014, pp. 1115-1149.

[21] G. shi, "Bayesian Classification," in *Data Mining and Knowledge Discovery for Geoscientists*, USA, Elsevier Inc, 2014, pp. 139-190.

[22] Q. He, "A review of clustering algorithms as applied to IR," University of Illinios, Urbana-Champaign, 1999.

[23] E. Alpaydin, Introduction to Machine Learning, second edition, The MIT Press, 2010.

[24] F. Kovacs, C. Legancy and A. Babos, "Cluster validity measurement techniques," in *6th international symposium of Hungarian researchers on computational intelligence*, USA, 2005.

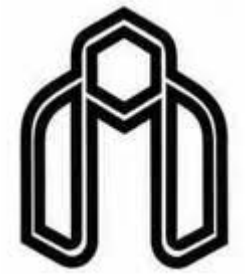
[25] J. Yick, B. Mukherjee and D. Ghosal, "Wireless sensor network survey," *Computer Networks,* vol. 52, pp. 2292-2330, 2008.

[26] "Data Sets for Data Fusion Experiments," [Online]. Available: http://lunadong.com/fusionDataSets.htm. [Accessed 1 Novemebr 2013].

Abstract:

In this thesis Data fusion in uncertainty condition defined as "combining several uncertainty sources to make an effective representation for human or machine by ability to handle uncertainty and inconsistency". The most important aspect of this thesis is identifying and dealing with uncertainty and inconsistency in unknown systems. Most previous data fusion methods such as Kalman filtering are dependent on the system behavior. Such dependency does not allow us to easily deal with unknown systems.in systems with behavior model, data fusion can be applied on two sources. But in unknown systems data fusion needs many sources. The proposed algorithm consists of a new clustering technique, neural network and finally new update prediction rules for the predictor. In general, when the sources contain uncertainty and inconsistency, data fusion may fail. The proposed method in this thesis can recognize and then remove the inconsistent data points, thus it presents reliable results. The experimental results on both synthetic and real data (weather forecast) confirm the effectiveness of the proposed approach. Finally we examined our proposed method using Meteorological data. The experimental results show strength of the proposed method and its ability to cope with uncertain and inconsistent sources.

Keywords: data fusion, uncertainty, clustering, neural network, weather prediction



Shahrood University of Technology

Faculty of Computer Engineering and IT

An efficient algorithm for data fusion under uncertainty conditions

Mohammad Mahdi Alyan Nezhadi

Supervisor:

Dr. Ali Akbar Pouyan

Advisor:

Dr. Vahid Abolghasemi

2014