





دانشگاه آزاد اسلامی
واحد دولت آباد
دانشکده فنی مهندسی

پایان نامه برای دریافت درجه کارشناسی ارشد در رشته کامپیوتر
گرایش نرم افزار

عنوان

ارائه یک طبقه‌بند نروفازی عمیق برای داده‌های پزشکی
نامتوازن چند کلاسه



معاونت پژوهش و فن آوری

به نام خدا

مشور اخلاق پژوهش

بیاباری از خداوند بجان و اعتقاد به اینکه عالم محضر خداست و همواره ناظر بر اعمال انسان و به منظور پادداشت مقام بلند دانش و پژوهش و نظریه اهمیت جایگاه دانشگاه در اعتلای فرهنگ و تمدن بشری و دانشجویان و اعضاء هیئت علمی و احد های دانشگاه آزاد اسلامی متعهد می گردیم اصول زیر را در انجام فعالیت های پژوهشی مد نظر قرار داده و از آن تخطی نکنیم:

- ۱- اصل حقیقت جویی: تلاش در راستای پی جویی حقیقت و وفاداری به آن و دوری از حرکت به پنهان سازی حقیقت
- ۲- اصل رعایت حقوق: التزام به رعایت کامل حقوق پژوهشگران و پژوهشگران (انسان، حیوان و نبات) و سایر صاحبان حق
- ۳- اصل مالکیت مادی و معنوی: تعهد به رعایت کامل حقوق مادی و معنوی دانشگاه و کلیه بکاران پژوهش
- ۴- اصل منافع ملی: تعهد به رعایت مصالح ملی و در نظر داشتن پیشبرد و توسعه کشور در کلیه مراحل پژوهش
- ۵- اصل رعایت انصاف و امانت: تعهد به اجتناب از حرکت به جانب داری غیر علمی و حفاظت از اموال، تجهیزات و منابع در اختیار
- ۶- اصل رازداری تعهد به صیانت از اسرار و اطلاعات محرمانه افراد سازمان ها و کشور و کلیه افراد و استاد های مرتبط با تحقیق
- ۷- اصل احترام: تعهد به رعایت حریم ها و حرمت ها در انجام تحقیقات و رعایت جانب تقد و خودداری از حرکت به حرمت شکنی
- ۸- اصل ترویج: تعهد به رواج دانش و اشاعه نتایج تحقیقات و انتقال آن به بکاران علمی و دانشجویان به غیر از مواردی که منع قانونی دارد
- ۹- اصل برانست: التزام به برانست جویی از حرکت به رفتار غیر حرفه ای و اعلام موضع نسبت به کسانی که حوزه علم و پژوهش را به شبه های غیر علمی می آلاینند

تعهدنامه اصالت رساله یا پایان نامه

اینجانب محسن قمی دانش آموخته مقطع کارشناسی ارشد در رشته مهندسی کامپیوتر
که در تاریخ..... از پایان نامه / رساله خود باعنوان.....

..... با کسب نمره و درجه دفاع نموده

ام بدینوسیله متعهد می شوم:

(۱) این پایان نامه / رساله حاصل تحقیق و پژوهش انجام شده توسط اینجانب بوده و در مواردی که از دستاوردهای علمی و پژوهشی دیگران (اعم از پایان نامه، کتاب، مقاله و ...) استفاده نموده ام، مطابق ضوابط و رویه موجود، نام منبع مورد استفاده و سایر مشخصات آن را در فهرست مربوطه ذکر و درج کرده ام.

(۲) این پایان نامه / رساله قبلاً " برای دریافت هیچ مدرک تحصیلی (هم سطح، پائین تر یا بالاتر) در سایر دانشگاه ها و مؤسسات آموزشی عالی ارائه نشده است .

(۳) چنانچه بعد از فراغت از تحصیل، قصد استفاده و هرگونه بهره برداری اعم از چاپ کتاب، ثبت اختراع و ... از این پایان نامه داشته باشم، از حوزه معاونت پژوهشی واحد مجوزهای مربوطه را اخذ نمایم.

(۴) چنانچه در هر مقطع زمانی خلاف موارد فوق ثابت شود، عواقب ناشی از آن را می پذیرم و واحد دانشگاهی مجاز است با اینجانب مطابق ضوابط و مقررات رفتار نموده و در صورت ابطال مدرک تحصیلی ام هیچ گونه ادعایی نخواهم داشت.

نام و نام خانوادگی:

تاریخ و امضاء:

فهرست مطالب

عنوان	صفحه
چکیده.....	۱
فصل اول کلیات تحقیق	۲
۱-۱- مقدمه	۲
۱-۲- بیان مسئله	۴
۱-۳- اهمیت تحقیق	۵
۱-۴- اهداف پایان نامه	۶
۱-۵- نوآوری پایان نامه	۶
۱-۶- ساختار پایان نامه	۷
فصل دوم مبانی و پیشینه تحقیق	۹
۱-۲- داده‌های نامتوازن	۹
۱-۱-۲- روش‌های سطح داده	۱۰
۲-۱-۲- روش‌های سطح الگوریتم	۱۶
۳-۱-۲- روش‌های حساس به هزینه	۱۷
۲-۲- هوش مصنوعی در تشخیص پزشکی	۱۸
۲-۳- داده کاوی	۲۲
۲-۴- دسته‌بندی	۲۳
۲-۵- مدل‌هایی براساس هوش مصنوعی	۲۵
۱-۲-۵- سیستم استنتاج عصبی-فازی تطبیقی	۲۶
۲-۲-۵- سیستم استنتاج فازی	۲۷
۳-۲-۵- شبکه تطبیقی	۲۸
۴-۲-۵- معماری سیستم استنتاج عصبی-فازی تطبیقی	۲۸
۵-۲-۵- الگوریتم یادگیری ترکیبی	۳۱

۳۲.....	۲-۶- الگوریتم بهینه‌سازی ازدحام ذرات
۳۳.....	۲-۷- کارهای مرتبط
۳۷.....	فصل سوم روش پیشنهادی
۳۷.....	۳-۱- مراحل اجرای روش پیشنهادی
۳۹.....	۳-۱-۱- مجموعه داده
۳۹.....	۳-۱-۲- پیش پردازش و آماده‌سازی
۴۱.....	۳-۱-۳- اعمال الگوریتم
۵۱.....	۳-۱-۴- آزمایش مدل
۵۱.....	۳-۱-۵- ارزیابی کارایی مدل
۵۳.....	فصل چهارم شبیه‌سازی و ارزیابی نتایج
۵۳.....	۴-۱- معیارهای ارزیابی
۵۵.....	۴-۲- عملکرد الگوریتم استنتاج عصبی فازی تطبیقی
۵۶.....	۴-۳- الگوریتم استنتاجی عصبی فازی بهینه شده
۵۸.....	۴-۴- تجزیه و تحلیل عملکرد مدل‌ها
۶۰.....	۴-۵- مقایسه نتایج
۶۱.....	فصل پنجم نتیجه‌گیری و پیشنهادات آتی
۶۲.....	۵-۲- پیشنهادات آتی

۱۸.....	شکل (۲-۱): انواع تکنیک‌های یادگیری ماشین
۲۲.....	شکل (۲-۲): مراحل فرآیند کشف دانش
۲۴.....	شکل (۲-۳): ساختار دسته‌بندی
۲۶.....	شکل (۲-۴): سیستم استنتاج فازی
۲۷.....	شکل (۲-۵): شبکه تطبیقی
۳۶.....	شکل (۳-۱): مراحل اجرای روش پیشنهادی
۳۹.....	شکل (۳-۲): ساختار سیستم استنتاج عصبی فازی تطبیقی
۴۰.....	شکل (۳-۳): سه پارامتر در تابع عضویت زنگوله‌ای (a,b,c)
۴۲.....	شکل (۳-۴): دیاگرام جریان پایه‌ای محاسبات در ANFIS
۴۵.....	شکل (۳-۵): بروزرسانی مکانیسم موقعیت PSO
۴۶.....	شکل (۳-۶): نموداری از ترتیب PSO و ANFIS
۴۸.....	شکل (۳-۷): بلوک دیاگرام تشخیص ANFIS-PSO برای طبقه‌بندی سرطان
۵۱.....	شکل (۴-۱): عملکرد مدل عصبی-فازی در مرحله آموزش مدل
۵۲.....	شکل (۴-۲): عملکرد مدل عصبی-فازی در مرحله آزمایش مدل
۵۳.....	شکل (۴-۳): عملکرد مدل عصبی-فازی بهینه شده در مرحله آموزش مدل
۵۳.....	شکل (۴-۴): عملکرد مدل عصبی-فازی بهینه شده در مرحله آزمایش مدل
۵۵.....	شکل (۴-۵): مقایسه خطای الگوریتم ANFIS و ANFIS-PSO در مرحله آموزش
۵۵.....	شکل (۴-۶): مقایسه خطای الگوریتم ANFIS و ANFIS-PSO در مرحله آموزش
۵۶.....	شکل (۴-۷): مقایسه روش پیشنهادی با کارهای پیشین

چکیده

درون سیستم‌های مراقبت از سلامت انبوهی از داده‌های بیماران ذخیره شده است. به کمک کشف دانش و داده‌کاوی، می‌توان به دانش مفید، ارتباطات و الگوهای مخفی درون داده‌ها دست یافت. دانش کشف‌شده می‌تواند توسط مدیران مراقبت از سلامت و پزشکان مورد استفاده قرار گیرد تا کیفیت خدمات، بهبود یابد و از طرفی نیز تعداد خطاهای پزشکی را بتوان کاهش داد. از آنجایی که استفاده از تنها یک الگوریتم داده‌کاوی، جهت تشخیص و یا پیش‌گویی بیماری‌ها کاری دشوار است، بنابراین به منظور بهبود طبقه‌بندی داده‌های نامتوازن، از شبکه عصبی فازی استفاده می‌شود. ایده اصلی این گونه شبکه‌ها از شیوه کارکرد سیستم عصبی زیستی برای پردازش داده‌ها و اطلاعات به منظور یادگیری و ایجاد دانش الهام گرفته شده است. این سیستم از شمار زیادی عناصر پردازشی فوق‌العاده بهم پیوسته به نام نورون تشکیل شده است که برای حل یک مساله با هم هماهنگ عمل می‌کنند. در این سیستم قسمت شبکه عصبی برای یادگیری و طبقه‌بندی توانایی‌ها و نیز پیوند و اصلاح الگو به کار می‌رود. قسمت شبکه عصبی به طور خودکار، قواعد منطق فازی و توابع عضویت را در طول دوره تناوب یادگیری ایجاد می‌کند. چند روش برای آموزش شبکه‌های عصبی وجود دارد که در این تحقیق از روش ترکیبی استفاده می‌شود. همچنین روش بهینه‌سازی ازدحام ذرات در این پایان‌نامه برای کمک به تنظیم پارامترهای توابع عضویت روش عصبی فازی استفاده شده است. مزیت اصلی روش بهینه‌سازی ازدحام ذرات، روش مساعد آن از محاسبه در یک توپولوژی شبکه از اندازه معین، است. پیاده‌سازی این سیستم در محیط متلب با مجموعه داده ویسکانسین سرطان سینه انجام شد و دارای دقت ۹۸٫۸٪ می‌باشد.

واژگان کلیدی: سیستم استنتاج عصبی فازی تطبیقی، بهینه‌سازی ازدحام ذرات، طبقه‌بندی، داده‌های نامتوازن پزشکی

فصل اول

کلیات تحقیق

۱-۱- مقدمه

یادگیری طبقه‌بندها از روی داده‌های نامتوازن یکی از مهمترین موضوعات رایج در زمینه‌ی مسائل مربوط به طبقه‌بندی است. زمانی که تعداد نمونه‌های مربوط به یک دسته نسبت به دسته‌ی دیگر بسیار کم باشد منجر به کاهش صحت و کارایی سیستم خواهد شد، چرا که ماشین‌های یادگیر با آموزش از روی داده‌های نامتوازن به سمت دسته‌ای که دارای نمونه‌های بیشتری باشد، گرایش پیدا خواهند کرد. عدم توازن در بسیاری از پایگاه داده‌ها وجود دارد، که هدف اصلی تشخیص نمونه‌هایی است که کم ولی دارای اهمیت بالایی هستند [۱].

روش‌های مختلف داده‌کاوی می‌تواند به پیش‌گویی دقیق بیماری‌ها کمک کند و از طرفی نیز باعث کاهش هزینه‌های اضافی که مثلاً ممکن است در اثر تجویز آزمایش‌های نامربوط بالینی ایجاد می‌شود، خواهد شد. همچنین در کاهش خطای تشخیص و پیش‌گویی اثر می‌گذارد و کیفیت خدمات نیز بهبود خواهد یافت. برخی از روش‌های داده‌کاوی که برای داده‌های پزشکی به کار برده می‌شوند عبارتند از: شبکه‌های عصبی، درخت‌های تصمیم، یافتن قانون انجمنی، شبکه‌های بیزی، ماشین بردار پشتیبان، خوشه‌بندی و ... هر کدام از این روش‌ها، بستگی به نوع کاربردها سودمند خواهند بود [۲]. اما استفاده از تنها یک الگوریتم داده‌کاوی، جهت تشخیص و یا پیش‌گویی تمام بیماری‌ها کاری دشوار است. بعضی از الگوریتم‌ها برای برخی کارها بهتر از دیگری هستند، اما هنگامی که از ترکیبی از مزایای برخی الگوریتم‌ها استفاده کنیم، نتایج بهتری در کارایی خواهیم داشت که ما

در این پایان‌نامه نیز به بررسی آن می‌پردازیم تا بتوانیم بهترین نوع روش داده‌کاوی را از لحاظ کارایی‌هایی چون دقت، حساسیت و ویژگی که در پیش‌گویی بیماری‌ها مرسوم هستند، معرفی کنیم. تحقیق روی داده‌ها و سردرگمی حاصل از آن مشکلی است که مانع رسیدن به نتایج قابل توجه می‌شود چرا که باید از دانش درون داده‌ها به نحو مناسب استفاده کرد.

در واقع داده‌کاوی، پاسخی به این نیاز سازمان‌های مراقبت از سلامت بوده است. هرچقدر که حجم داده‌ها نیز بیشتر باشد و روابط بین آن‌ها نیز پیچیده‌تر باشد، دسترسی به اطلاعات نهفته درون داده‌ها نیز مشکل‌تر می‌شود [۳]. در الگوریتم‌های استاندارد یادگیری معمولاً فرض بر این است که توزیع رده‌ها متوازن یا نزدیک به متوازن باشند. همچنین در حالت کلی هزینه طبقه‌بندی اشتباه نیز یکسان فرض می‌شود. بنابراین، هنگامی که مجموعه‌های داده نامتوازن است این الگوریتم‌ها در نمایش مشخصات توزیع داده‌ها به صورت صحیح دچار مشکل می‌شوند؛ به تعبیری این الگوریتم‌ها متمایل به رده‌های با فراوانی بیشتر می‌شوند و در نتیجه صحت غیر قابل قبولی را در بین رده‌های داده فراهم می‌کنند. داده‌های نامتوازن عبارت است از هر مجموعه داده‌ای که نشان دهنده یک توزیع نامتوازن در بین رده‌های خود باشند به طوری که میزان نامتوازن بودن توزیع بسیار زیاد باشد. این نوع عدم توازن را عدم توازن بین رده‌ها می‌نامند. البته توزیع نامتوازن لزوماً بین دو رده نبوده و ممکن است بین چند رده وجود داشته باشد. در بسیاری از داده‌های واقعی توزیع داده‌ها به صورت نامتوازن است و بر همین اساس لازم است تا الگوریتم‌های یادگیری برای استفاده از دانش موجود در این داده‌ها اصلاح شوند [۴]. از جمله این مجموعه داده‌ها که دارای توزیع نامتوازن هستند، می‌توان داده‌های بیماران مبتلا به سرطان را مثال زد. این داده‌ها اغلب با دو رده مثبت که مبتلا به سرطان را نشان می‌دهد و منفی که سلامت بیمار را نشان می‌دهد، نشان داده می‌شوند. همان‌طور که انتظار می‌رود تعداد افراد سالم بسیار بیشتر از افراد مبتلا به سرطان است. بنابراین نوعی رده‌بندی مورد نیاز است که برای چنین مجموعه داده‌ای میزان دقت پیش‌بینی مناسب و متوازن برای هر دو رده اقلیت و اکثریت به دست آورد. از آن جایی که می‌دانیم در پزشکی تشخیص یک بیمار سرطانی به عنوان سالم و البته به طور مشابه، تشخیص یک فرد سالم به عنوان یک فرد بیمار غیرقابل قبول است. طبقه‌بندهای مورد استفاده باید بتوانند صحت بالایی را برای رده اقلیت فراهم کرده و در عین حال صحت رده اکثریت را نیز چندان تحت تاثیر قرار ندهند.

هدف از این پایان‌نامه بررسی روش‌های حل مشکل مواجهه با داده‌های نامتوازن پزشکی در زمینه‌ی طبقه‌بندی است. با افزایش حجم داده‌ها و ظهور داده‌های نامتوازن در دهه‌ی اخیر، نیاز برای طبقه‌بندی هرچه بهتر این داده‌ها احساس می‌شود. از طرف دیگر سیستم‌های مبتنی بر یادگیری جمعی کارایی بالایی خود را در زمینه‌ی مسائل طبقه‌بندی به‌خوبی نشان داده‌اند. دسته‌بندی داده‌های نامتوازن با استفاده از این سیستم‌های طبقه‌بندی و نظریه‌ی نرو- فازی می‌تواند دقت عملکرد بهتری را در این زمینه داشته باشد.

۱-۲- بیان مسئله

برای طبقه‌بندی داده‌های نامتوازن چالش‌هایی وجود دارد که الگوریتم‌های طبقه‌بندی معمولی، قادر به غلبه بر آن‌ها نیستند. چالش نخست این است که طبقه‌بند‌های معمولی، برای عملکرد صحیح، احتیاج دارند تا با داده‌های با توزیع کلاسی تقریباً یکسان آموزش داده شوند. در نتیجه این طبقه‌بندها در مواجهه با داده‌های نامتوازن دچار مشکل می‌شوند. در این‌گونه مسائل، طبقه‌بندها تمایل شدیدی برای تمرکز بر روی داده‌های کلاس اکثریت پیدا می‌کنند. این تمایل شدید منجر به چشم‌پوشی از داده‌های کلاس اقلیت یا در نظر گرفتن این داده‌ها به‌عنوان نقاط نویزی در زمان آموزش می‌شود.

چالش دوم این است که با چشم‌پوشی نمودن طبقه‌بند از داده‌های اقلیت، مدل آموزش داده‌شده صرفاً قدرت تشخیص داده‌های کلاس اکثریت را پیدا می‌کند. با توجه به اینکه داده‌های این کلاس چندین برابر داده‌های کلاس دیگر است، سیستم طبقه‌بندی در مجموع از دقت بسیار بالایی برخوردار می‌شود؛ اما در طبقه‌بندی داده‌های اقلیت به‌شدت ضعیف عمل خواهد کرد [۵]. فائق آمدن بر این مشکل نیازمند استفاده از روش‌ها و الگوریتم‌هایی است که با وجود تعداد کم نمونه‌های کلاس اقلیت، نه‌تنها آن‌ها را برای آموزش در نظر بگیرد، بلکه تمرکز اصلی برای آموزش را نیز بر روی این داده‌ها قرار دهد. همان‌طور که در ادامه‌ی این پژوهش بیان خواهد شد، طبقه‌بندی داده‌های نامتوازن با چالش‌های زیادی روبه‌رو است که روش‌های ارائه شده در سال‌های اخیر هر کدام سعی در غلبه کردن بر این چالش‌ها به نحوی داشته‌اند. اما نکته‌ی حائز اهمیت این است که روش‌های موجود اکثراً به متوازن نمودن مجموعه‌ی داده اهمیت می‌دهند و به افزایش کیفیت فضای ویژگی برای بهبود دقت طبقه‌بندی توجهی نمی‌کنند. علاوه بر این، جز تعدادی انگشت شمار از روش‌های مرسوم،

روش‌های دیگر به استفاده از قدرت بالای مجموعه‌های فازی در طبقه‌بندی توجه نکرده‌اند. لذا با توجه به سوابق درخشان استفاده از مجموعه‌های فازی در طبقه‌بندی فازی، استفاده‌ی هدفمندتر از این مجموعه‌ها برای افزایش دقت طبقه‌بندی داده‌های نامتوازن منطقی به نظر می‌آید.

در این پژوهش به دنبال آن هستیم تا با استفاده از روش‌های نرو- فازی و روش‌های مختلف برخورد با مجموعه‌های نامتوازن، طبقه‌بندی این داده‌ها را بهبود بخشیم. روش‌های نرو- فازی می‌توانند با استفاده از عدم قطعیت و انعطاف ذاتی که دارند، باعث افزایش دقت طبقه‌بندی به‌ویژه داده‌های کلاس اقلیت شوند.

۱-۳- اهمیت تحقیق

مجموعه داده‌های نامتوازن سهم زیادی از مجموعه داده‌های جهان واقعی را تشکیل می‌دهند. در این مجموعه داده‌ها نمونه‌های کلاس حداقلی نسبت به نمونه‌های کلاس حداکثری اهمیت بالاتری دارند؛ از این‌رو طبقه‌بندی صحیح این نمونه‌ها یکی از مسایل اساسی در داده کاوی به شمار می‌آید [۵].

چنانچه به عنوان نمونه مجموعه داده‌های نامتوازن را مجموعه اطلاعات افراد بیمار فرض کنیم دسته‌ای از این افراد مبتلا به سرطان (کلاس حداقلی) و دسته‌ای دیگر سالم هستند. همان‌گونه که مشخص است افراد سرطانی تعداد بسیار کمتری در کل مجموعه افراد هستند؛ اما اهمیت بیشتری دارند. در سیستم طبقه‌بندی هدف تعیین نوع کلاس یک نمونه جدید است.

مسئله طبقه‌بندی داده‌های نامتوازن بسیار حائز اهمیت می‌باشد؛ زیرا به طور ضمنی در اکثر کاربردهای دنیای واقعی مشاهده می‌شود؛ مانند پیش‌بینی ورشکستگی، تشخیص کلاهبرداری، تشخیص چهره، تجارت الکترونیک، تشخیص نفوذ و مخصوصاً در تشخیص‌های پزشکی.

۱-۴- اهداف پایان نامه

هدف از این پایان نامه طبقه‌بندی داده‌های نامتوازن مربوط سرطان سینه می‌باشد. مجموعه داده‌های نامتوازن یکی از چالش‌های پیش روی طراحی سیستم‌های پزشکیار برای طبقه‌بندی و تعیین وضعیت بیمار محسوب می‌شود که در این پژوهش از روش‌های سطح الگوریتم برای حل آن استفاده می‌شود. از جمله اهداف این پژوهش می‌توان به موارد زیر اشاره کرد:

- افزایش دقت طبقه‌بندی داده‌های نامتوازن با استفاده از الگوریتم نرو-فازی بهینه شده
- بهبود صحت در تشخیص بیماری سرطان سینه
- بهبود صحت و درستی الگوریتم پیشنهادی نسبت به سایر الگوریتم‌های موجود

۱-۵- نوآوری پایان نامه

در این پایان نامه برای طبقه‌بندی داده‌های نامتوازن سرطان، از ترکیب الگوریتم سیستم استنتاج عصبی-فازی تطبیقی با الگوریتم‌های فراابتکاری استفاده می‌گردد. به این ترتیب از الگوریتم‌های فرا ابتکاری برای بهبود کارایی ANFIS با تنظیم توابع عضویت و به حداقل رساندن خطا استفاده می‌کنیم. بنابراین، در این پژوهش با استفاده از روش فازی عصبی ویژگی‌های بیماری سرطان را مدل‌سازی می‌کنیم. در بین ترکیب‌های مختلف موجود در محاسبات، ترکیب منطق فازی با شبکه‌های عصبی در قالب سیستم‌های فازی عصبی از جلوه‌ی بیشتری برخوردار است. این سیستم‌ها نقش مهمی را در استنتاج قواعد از مشاهدات ایفا می‌کنند. یک روش کارا در این راستا توسط دکتر راجر جانگ در سال ۱۹۹۵، تحت عنوان روش فازی عصبی توسعه داده شده است. در چنین سیستم‌هایی بوسیله‌ی ورودی و خروجی‌های موجود و با کمک دانش بشری، در قالب قوانین اگر-آنگاه فازی برای ورودی دلخواه، خروجی مطلوب حاصل خواهد شد.

سیستم استنتاج فازی عصبی تطبیقی از ترکیب سامانه فازی با شبکه‌های عصبی تشکیل یافته است که مزایای هر دو روش را در بر می‌گیرد. به همین دلیل برای جداسازی و تفکیک ویژگی‌ها از سیستم استنتاج فازی عصبی تطبیقی و بهینه شده استفاده شد.

هرکدام از سیستم‌های فازی و شبکه‌های عصبی مصنوعی مزایا و نواقصی دارند. سیستم فازی قادر به استفاده از زبان بشری است و می‌تواند از تجربیات بشری و افراد متخصص و خبره استفاده کند، درحالی که قادر به یادگیری نیست. به عبارت دیگر با استفاده از داده‌های مشاهده‌ای نمی‌توان سیستم فازی را آموزش داد. اما شبکه‌های عصبی با استفاده از مجموعه داده‌ها، قابلیت خودآموزی دارند. در عین حال شبکه‌های عصبی غیرصریح هستند و قادر به استفاده از زبان بشری نیستند.

۱-۶- ساختار پایان نامه

ساختار کلی این پایان‌نامه این‌گونه خواهد بود:

- در فصل دوم به مبانی و معرفی پژوهش‌ها و کارهای انجام‌شده در زمینه‌ی طبقه‌بندی داده‌های نامتوازن پرداخته خواهد شد.
- روش پیشنهادی این پژوهش در فصل سوم به تفصیل بیان خواهد شد.
- در فصل چهارم نتایج و عملکرد روش پیشنهادی در مقایسه با روش‌های مرسوم و شناخته‌شده در این زمینه ارائه می‌شوند.
- در نهایت نتیجه‌گیری و پیشنهادات این پایان‌نامه در فصل پنجم انجام می‌گیرد.

فصل دوم

مبانی و پیشینه تحقیق

یکی از مسائل مهم در زمینه داده کاوی، مسأله طبقه‌بندی مجموعه داده‌های نامتوازن است. اصطلاح مجموعه داده نامتوازن، عموماً به مجموعه داده‌ای گفته می‌شود که تعداد نمونه‌ها در کلاس‌های گوناگون، اختلاف بسیاری داشته باشند. در این فصل از پایان‌نامه، ابتدا به حل مسئله طبقه‌بندی داده‌های نامتوازن پرداخته شده، سپس الگوریتم‌های استفاده شده در این پایان‌نامه معرفی می‌شوند و در آخر هم کارهای مرتبط شرح داده شده است.

۲-۱- داده‌های نامتوازن

برای حل مسئله طبقه‌بندی داده‌های نامتوازن، الگوریتم‌ها و روش‌های مختلفی ارائه شده است که این روش‌ها را می‌توان در سه دسته‌ی کلی جای داد [۶]:

۱- روش‌های سطح داده

۲- روش‌های سطح الگوریتم

۳- روش‌های حساس به هزینه