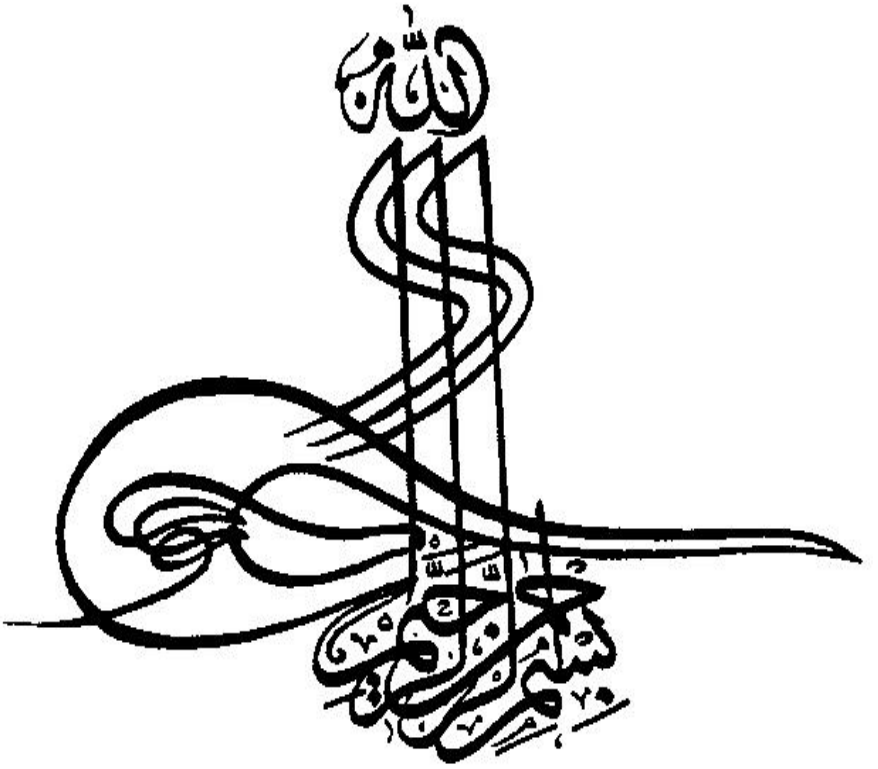

دانشگاه کردستان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان:
بهبود کارایی دسته بندی متن بر مبنای ویژگیها و متون دسته‌بندی شده مشابه

پژوهشگر:

جلیل بیات

استاد راهنما:

دکترفردین اخلاقیان طاب

استاد مشاور:

دکتر پرهام مرادی

مهر ماه ۱۳۹۳

کلیه حقوق مادی و معنوی مترتب بر نتایـج مطالعات،

ابتکارات و نوآوری های ناشی از تحقیق موضوع

این پایان‌نامه (رسـاله) متعلق به دانشـگاه کردستان است.

تعهد نامه

اینجانب جلیل بیات دانشجوی کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی دانشگاه کردستان، دانشکده فنی و مهندسی گروه مهندسی کامپیوتر و فناوری اطلاعات تعهد می‌نماییم که محتوای این پایان نامه نتیجه تلاش و تحقیقات خود بوده و از جایی کپی برداری نشده و به پایان رسانیدن آن نتیجه تلاش و مطالعات مستمر اینجانب و راهنمایی و مشاوره اساتید بوده است.

با تقدیم احترام

جلیل بیات

۱۳۹۳/۶/۲

بسمه تعالی

* تعهد نامه دانشجویان تحصیلات تکمیلی دانشگاه کردستان در انجام پایان نامه *

(لازم است به عنوان صفحه اول پروپوزال و به عنوان چهارمین برگ پایان نامه و پس از صفحه مشخصات پایان نامه بوده و به دقت مطالعه و امضا شود)

اینجانب دانشجوی مقطع رشته متعهد میشوم:

۱- صداقت، امانتداری و بی طرفی را در انجام پژوهش و انتشار نتایج حاصل از آن رعایت نمایم.

۲- در نگارش نتیجه پژوهش های حاصل از موضوع پایان نامه، از باز نویسی نوشته های دیگران بدون ذکر منبع، بازی با الفاظ، زیاده نویسی، کلی گویی و جزم اندیشی و تصرف گرایی پرهیز نمایم و نتایج پژوهشی خود را در موعد مقرر و با اطلاع استاد راهنما منتشر نمایم.

۳- تمامی یافته های مستخرج از پایان نامه متعلق به دانشگاه کردستان بوده و لازم است در کلیه مقالات مستخرج از آنها نام دانشگاه کردستان را تحت عنوان ((دانشجوی دانشگاه کردستان)) یا ((دانش آموخته دانشگاه کردستان)) ذکر نمایم.

۴- در انتشار مقالات نام استاد (استادان) راهنما و استاد (استادان) مشاور را در لیست مولفین مقاله ذکر نمایم و از آوردن اسامی افرادی که نقش موثری در انجام پژوهش نداشته اند، جداً خودداری نمایم.

۵- در بخش سپاسگزاری مقاله، از تمامی افراد و سازمانهایی که در اجرای پژوهش مساعدتی مبذول داشته اند با ذکر نوع مشارکت تشکر و قدر دانی نمایم.

۶- از انتشار همپوشان یا ارسال همزمان یک مقاله به چند مجله وبا ارسال مجدد مقاله چاپ شده به مجلات دیگر خودداری نمایم.

۷- در صورت عدم رعایت موارد مذکور، دانشگاه کردستان مجاز خواهد بود تا برابر مقررات اقدام نماید.

امضاء و اثر انگشت دانشجو

دستورالعمل نحوه برخورد با موارد تخطی دانشجویان تحصیلات تکمیلی در هنگام انتشار نتایج پژوهش

۱- در موارد زیر دانشگاه کردستان با مجله مربوطه مکاتبه و درخواست خارج نمودن مقاله را نموده و موضوع را به محل کار یا تحصیل بعدی دانشجو اطلاع خواهد داد.

الف – چاپ مقاله بدون اطلاع و تأیید استادان راهنما،

ب – چاپ نتایج حاصل از پژوهش های انجام شده در دانشگاه کردستان بدون ذکر نام دانشگاه

۲- در صورت احراز تخلف ازسایر موارد درج شده در تعهد نامه دانشجویی، دانشگاه ضمن مکاتبه با مجله مربوطه، حسب مورد تصمیم گیری خواهد نمود.

دانشگاه کردستان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر و فناوری اطلاعات

پایان نامه کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی

عنوان:
بهبود کارایی دسته بندی متن بر مبنای ویژگیها و متون دسته‌بندی شده مشابه

پژوهشگر:

جلیل بیات

در تاریخ / / ۱۳ توسط کمیته تخصصی وهیات داوران زیر مورد بررسی قرار گرفت و با نمره ..... و درجه ..... به تصویب رسید.

هیات داوران	نام و نام خانوادگی	مرتبه علمی	امضاء
۱ـ استاد راهنما	دکتر فردین اخلاقیان طاب	استادیار	
۲ـ استاد مشاور	دکتر پرهام مرادی	استادیار	
۳ـ استاد داور خارجی	دکتر محمد فتحی	استادیار	
۴ـ استاد داور داخلی	دکتر علیرضا عبدالله پوری	استادیار	

مهر و امضاء مدیر گروه مهر و امضاء معاون آموزشی و تحصیلات تکمیلی دانشکده

سپاسگذاری و قدردانی

و بعد از مدتها، پس از پیمودن راههای فراوان که با حضور شیرین **اساتید عزیزم**، با راهنماییها و دغدغه‌های فراوانشان، نگاههای **پدر و مادرم**، با چشمهای پر از برق شوق، و زیبایی حضور **خواهرم** در کنارم، که خستگی‌های این راه را به امید و روشنی راه تبدیل کرده و امیدوارم بتوانم در آینده‌ای نزدیک جوابگوی این همه محبت آنها باشم...

اکنون، با احترام فراوان برای این همه تلاش این عزیزان برای موفقیت من....

این پایان نامه را به **پدر و مادرم**، **اساتید عزیز**، **بویژه آقای دکتر فردین اخلاقیان** و**خواهر مهربانم** تقدیم می‌کنم.

امیدوارم قادر به درک زیباییهای وجودشان باشم.

با تشکر

جلیل بیات

چکیده

در یادگیری ماشین، داده‌های آموزشی نقش مهمی را در تعیین کارایی الگوریتم یادگیری ایفا می‌کنند. تعداد این داده‌های آموزشی در طول زمان افزایش می‌یابد وداده‌های جدیدی از راه می‌رسد. این داده‌های جدید، ممکن است دید جدیدی از داده‌های قبلی را ارائه دهند یا توزیع آماری داده‌ها را تغییر دهند. از این رو، فهمیدن اهمیت داده‌های جدید و اجازه به این داده‌ها، برای نقش داشتن در آموزش، به منظور افزایش کارایی سیستم یادگیر، کاری بس زیرکانه است. آموزش دوباره‌ی سیستم یادگیر با همه‌ی داده‌ها، و کنار گذاشتن همه‌ی تجربه‌ی یادگیری گذشته، یکی از راه‌حل‌ها برای این مسئله است ولی این روش دارای نقص‌هایی می‌باشد ازجمله اینکه این روش نا کارآمد است و همچنین، این روش قادر به نگه‌داری دانش قبلی نمی‌باشد.

در این پایان نامه، ما چهار روش ارائه نمودیم. در روش اول، ما سیستم یادگیری افزایشی برای دسته‌بندی متن ارائه نمودیم، که در این سیستم، برای هر دسته‌ی آموزشی از یک دسته‌بند ماشین بردار پشتیبان برای یادگیری آن استفاده نمودیم و سپس، برای دسته‌بند بعدی علاوه بر دسته‌ی آموزشی آن دسته، از بردارهای پشتیبان دسته‌های آموزشی قبلی به اضافه‌ی داده‌های غلط دسته بندی شده‌ی مجموعه‌های تصدیق دسته‌های آموزشی قبلی استفاده نمودیم. در روش دوم برای اینکه بتوانیم میزان معیارهای ارزیابی کارایی را افزایش و خطای دسته‌بندی را کاهش دهیم به جای استفاده از تنها یک دسته‌بند برای هر دسته‌ی آموزشی، از چند دسته‌بند برای هر دسته‌ی آموزشی استفاده نمودیم و همچنین، به جای در نظر گرفتن وزن یکسان به هر دسته‌بند برای تاثیر در ترکیب خروجی‌ها، از روش اول پایان‌نامه‌ی آقای علی دانش استفاده نمودیم که ایشان به ازای ترکیب هر دسته و دسته‌بند از وزن مستقلی استفاده نموده‌اند. در روش سوم، به منظور کامل کردن روش قبلی، برای هر دسته‌ی آموزشی از میان مجموعه‌ی دسته‌بندها، چندین دسته‌بند قابل قبول را انتخاب نمودیم هدف از ارائه این روش کنارگذاشتن دسته‌بندهای ضعیف و جایگزین کردن آنها با دسته‌بندهای قویی بود که با داده‌های آموزشی بیشتری آموزش دیده بودند بود که این امر اشاره به قابلیت خود تطبیقی سیستم پیشنهادی دارد. در روش چهارم با بسط روش سوم در هنگام گرفتن خروجی نهایی، نودهای ایجاد کننده‌ی نویز را حذف و نودهایی، برای افزایش کارایی دسته‌بندی اضافه نمودیم.

عملکرد روش‌های پیشنهادی با پنج روش دیگر مقایسه شد. نتایج آزمایشات، کارایی روش‌های پیشنهادی و بهبود کارایی دسته‌بندی متن را نشان می‌دهد.

کلمات کلیدی: دسته‌بندی متن، انتخاب ویژگی، یادگیری افزایشی، دسته‌بند، مرحله‌ی آموزشی، دسته‌ی آموزشی

فهرست مطالب

عنوان **صفحه**

فصل۱. مقدمه ۱
۱-۱-مقدمه ۱
۱-۲-بیان مساله تحقیق ۱
۱-۳-اهداف و رویکردها ۲
۱-۴-دستاوردهای پایان نامه ۳
۱-۵-ساختار پایان نامه ۴
فصل۲. پیشینه ۶
۲-۱-مقدمه ۶
۲-۲-نحوه‌ی بازنمایی اسناد بصورت بردارهای عددی ۷
۲-۲-۱-استخراج ویژگی ۹
۲-۳-دسته‌بندها ۱۲
۲-۳-۱-دسته‌بند ماشین بردار پشتیبان ۱۲
۲-۳-۲-دسته‌بند بیز ۱۵
۲-۳-۳-دسته‌بند درخت تصمیم ID3 ..... ۱۶
۲-۳-۴-دسته بند نزدیکترین همسایه ۲۰
۲-۳-۵-درخت تصمیم C4.5 ..... ۲۱
۲-۴-مروری بر روش‌های قبلی یادگیری افزایشی ۲۲
۲-۴-۱-سطح ۱ تطبیق: ۲۷
۲-۴-۲-سطح ۲ تطبیق: ترکیب دسته‌بندها ۲۷
۲-۴-۳-سطح ۳ از تطبیق: وزندهی به دسته‌بندها ۲۹
۲-۴-۴-یادگیری افزایشی براساس هرس ترکیب دسته‌بندها ۳۴
۲-۵-جمع بندی ۳۸
فصل۳. روش‌های پیشنهادی ۳۹
۳-۱-مقدمه ۳۹

۳-۲-روش اول: استفاده از داده‌های غلط دسته‌بندی شده و بردارهای پشتیبان دسته‌های آموزشی قبلی ۴۰
۳-۳-روش دوم: اضافه کردن تعداد دسته‌بندها برای هر دسته ی آموزشی و ترکیب دسته بندها براساس روش پیشنهادی آقای علی دانش ۴۳
۳-۴-جمع بندی..... ۴۶
فصل۴. اصلاح وتکامل روش‌های پیشنهادی ۴۸
۴-۱-مقدمه ۴۸
۴-۲-روش سوم: انتخاب چند دسته‌بند کارا از میان مجموعه‌ی دسته‌بندها و اصلاح دسته‌بندهای با کارایی پایین ۴۸
۴-۳-روش چهارم: حذف و اضافه کردن نود ۵۱
۴-۴-شمای کلی الگوریتم پیشنهادی..... ۵۳
۴-۵-جمع بندی..... ۵۷
فصل۵. نتایج و تحلیل زمایش ۵۹
۵-۱-مقدمه ۵۹
۵-۲-مجموعه داده‌های مورد استفاده ۵۹
۵-۲-۱-مجموعه‌ی داده‌ی20Newsgroups..... ۵۹
۵-۲-۲-مجموعه داده‌ی WebKB..... ۶۰
۵-۲-۳-مجموعه داده‌ایReuters-R8 ..... ۶۱
۵-۲-۴-معیارهای میکرو ۶۲
۵-۲-۵-معیارهای ماکرو ۶۳
۵-۳-مقایسه‌ی روش‌های پیشنهادی با یکدیگر ۶۴
۵-۳-۱-بررسی تاثیر افزایش تعداد دسته‌بندها در افزایش کارایی دسته‌بندی ۶۴
۵-۳-۲-بررسی تاثیر روش وزن‌دهی آقای علی دانش به دسته‌بندها در افزایش کارایی دسته‌بندی..... ۶۶
۵-۳-۳-تاثیر استفاده از روش اصلاح نود در بالابردن کارایی دسته‌بندی..... ۶۸
۵-۳-۴-تاثیر حذف و اضافه نمودن نود در بالا بردن کارایی دسته بندی ۷۰
۵-۴-مقایسه‌ی روش‌های پیشنهادی با سایر روش‌ها ۷۲
۵-۴-۱-مقایسه‌ی روش اول با سایر روش‌ها..... ۷۲

۵-۴-۲-مقایسه‌ی روش دوم با سایر روش‌ها ۷۴
۵-۴-۳-مقایسه‌ی روش سوم با سایر روش‌ها..... ۷۷
۵-۴-۴-مقایسه‌ی روش چهارم با سایر روش‌ها..... ۸۰
۵-۵-جمع‌بندی..... ۸۳
فصل۶. جمع بندی و پژوهش‌های آتی..... ۸۴
۶-۱-جمع بندی..... ۸۴
۶-۲-پژوهش‌های آتی..... ۸۶

فهرست جداول

عنوان **صفحه**

جدول ۵-۱: مشخصات مجموعه داده‌ای ۲۰ گروه خبری ۶۰

جدول ۵-۲: مشخصات مجموعه داده‌ای WebKb ۶۱

جدول ۵-۳: مشخصات مجموعه داده‌ایReuters-R8 ..... ۶۲

جدول ۵-۴: مقادیراولیه نگه دارنده ومقادیر افزایش آن در هر مرحله ۶۴

جدول ۵-۵: مقایسه‌ی روش پیشنهادی اول و روش پیشنهادی دوم برای داده‌هایReuters-R8 ... ۶۵

جدول ۵-۶: مقایسه‌ی روش پیشنهادی اول و روش پیشنهادی دوم برای داده‌هایWebKb ۶۵

جدول ۵-۷: مقایسه‌ی روش پیشنهادی اول و روش پیشنهادی دوم برای داده‌هایNewsgroup20 ..... ۶۵

جدول ۵-۸: مقایسه‌ی روش پیشنهادی دوم با ترکیب دسته بندها به صورت اکثریت آراء ساده و روش وزن‌دهی روش اول پایان‌نامه‌ی آقای علی دانش برای داده‌های Reuters-R8 ..... ۶۷

جدول ۵-۹: مقایسه‌ی روش پیشنهادی دوم با ترکیب دسته بندها به صورت اکثریت آراء ساده و روش وزن‌دهی روش اول پایان‌نامه‌ی آقای علی دانش برای داده‌های WebKb ۶۷

جدول ۵-۱۰: مقایسه‌ی روش پیشنهادی دوم با ترکیب دسته بندها به صورت اکثریت آراء ساده و روش وزن‌دهی روش اول پایان‌نامه‌ی آقای علی دانش برای داده‌های 20Newsgroup ۶۷

جدول ۵-۱۱: مقایسه‌ی روش پیشنهادی دوم با روش پیشنهادی سوم برای مجموعه داده‌ای Reuters-R8 ..... ۶۸

جدول ۵-۱۲: مقایسه‌ی روش پیشنهادی دوم با روش پیشنهادی سوم برای مجموعه داده‌ای WebKb ۶۹

جدول ۵-۱۳: مقایسه‌ی روش پیشنهادی دوم با روش پیشنهادی سوم برای مجموعه داده‌ای 20Newsgroup ۶۹

جدول ۵-۱۴: مقایسه‌ی روش پیشنهادی سوم-بخش اول و سوم-بخش دوم با روش پیشنهادی دوم برای مجموعه داده‌ای Reuters-R8 ..... ۷۰

جدول ۵-۱۵: مقایسه‌ی روش پیشنهادی سوم-بخش اول و سوم-بخش دوم با روش پیشنهادی دوم برای مجموعه داده‌ای WebKb ۷۱
جدول ۵-۱۶: مقایسه‌ی روش پیشنهادی چهارم-بخش اول و چهارم -بخش دوم با روش پیشنهادی سوم برای مجموعه داده‌ای 20Newsgroup ۷۱
جدول ۵-۱۷: مقایسه‌ی روش اول با سایر روش‌ها برای داده‌های Reuters-R8 ..... ۷۳
جدول ۵-۱۸: مقایسه‌ی روش اول با سایر روش‌ها برای داده‌های WebKb ۷۳
جدول ۵-۱۹: مقایسه‌ی روش اول با سایر روش‌ها برای داده‌های 20Newsgroup ۷۳
جدول ۵-۲۰: مقایسه‌ی روش دوم و سایر روش‌ها برای مجموعه داده‌ای Reuters-R8 ..... ۷۵
جدول ۵-۲۱: مقایسه‌ی روش دوم با سایر روش‌ها برای مجموعه داده‌ای WebKb ۷۶
جدول ۵-۲۲: مقایسه‌ی روش دوم با سایر روش‌ها برای مجموعه داده‌ای 20Newsgroup ۷۷
جدول ۵-۲۳: نتایج روش سوم وسایر روش‌ها برای مجموعه داده‌ای Reuters-R8 ..... ۷۸
جدول ۵-۲۴: نتایج معیارهای ارزیابی روش سوم وسایر روش‌ها برای مجموعه داده‌ای WebKb ۷۹
جدول ۵-۲۵: نتایج معیارهای ارزیابی روش سوم وسایر روش‌ها برای مجموعه داده‌ای 20Newsgroup ۸۰
جدول ۵-۲۶: نتایج معیارهای ارزیابی روش چهارم وسایر روش‌ها برای مجموعه داده‌ای..... ۸۱
جدول ۵-۲۷: نتایج معیارهای ارزیابی روش چهارم وسایر روش‌ها برای مجموعه داده‌ای WebKb ۸۲
جدول ۵-۲۸: نتایج معیارهای ارزیابی روش چهارم وسایر روش‌ها برای مجموعه داده‌ای 20NewsGroup ۸۲

فهرست شکل‌ها

عنوان **صفحه**

شکل ۲-۱: نحوه پیش پردازش متن ۸
شکل ۲-۲: مثالی از یک درخت تصمیم [۳۰] ۱۷
شکل ۲-۳: شبه‌کد درخت تصمیم ID3 [۳۰] ۱۸
شکل ۲-۴: مثالی از روش نزدیکترین همسایه (K-6) [۱۲] ۲۰
شکل ۲-۵: شبه کد درخت تصمیم C4.5 [۳۴] ۲۲
شکل ۲-۶: دسته‌بندی خودکار متن با استفاده از داده‌های بدون برچسب [۳۵] ۲۳
شکل ۲-۷: شبه کد یادگیری با استفاده از داده‌های بدون برچسب [۳۵] ۲۴
شکل ۲-۸: مدل یادگیری افزایشی با استفاده از ترکیب دسته‌بندها [۱] ۲۵
شکل ۲-۹: شبه کد یادگیری افزایشی با استفاده از ترکیب دسته‌بندها [۱] ۲۵
شکل ۲-۱۰: تطبیق چند سطحی [۳۶] ۲۶
شکل ۲-۱۱: ترکیب یادگیرنده‌های برخط [۳۶] ۲۸
شکل ۲-۱۲: شبه کد وزن‌دهی [۳۶] ۲۹
شکل ۲-۱۳: شبه کد ترکیب اکثریت آراء وزن دار [۳۶] ۳۰
شکل ۲-۱۴: شمای کلی الگوریتم ADAIN [۳۷] ۳۱
شکل ۲-۱۵: شبه کد ADAIN [۳۷] ۳۲
شکل ۲-۱۶: تابع نگاشت با استفاده از MLPs [۳۷] ۳۳
شکل ۲-۱۷: شبه کد هرس [۳۷] ۳۴
شکل ۲-۱۸: شبه‌کد ++PBagging [۳] ۳۵
شکل ۲-۱۹: دسته بندی SV-Incremental [۳۹] ۳۶
شکل ۲-۲۰: شبه‌کد SV-Incremental [۳۹] ۳۷
شکل ۳-۱: استفاده از بردارهای پشتیبان در بهبود دسته‌بندی [۳۹] ۴۱
شکل ۳-۲: شمای کلی روش اول ۴۲

شکل ۳-۳: شمای کلی روش دوم ۴۵
شکل ۴-۱: شمای کلی روش سوم ۵۰
شکل ۴-۲: الگوریتم حذف نود ایجاد کننده‌ی نویز ۵۲
شکل ۴-۳: فلوچارت کلی الگوریتم یادگیری افزایشی پیشنهادی برای دسته‌بندی متن ۵۶

فصل ۱. مقدمه

۱-۱- مقدمه

مغز انسان دارای توانایی بالایی، در یادگیری افزایشی است. بنابراین شبیه سازی توانایی مغز انسان در یادگیری افزایشی یک از مسائل چالش برانگیز در یادگیری ماشین است. در کاربردهای دنیای واقعی، سه زمینه وجود دارد که نیازمند یادگیری افزایشی است: ۱)امکان جمع‌آوری داده‌های آموزشی بصورت یکجا امکان پذیر نیست به عنوان مثال در حیطه‌ی پزشکی ۲) در برخی از مسائل جهان واقعی به محض اینکه دسته‌ی آموزشی‌ای رسید ما نیازمند یادگیری هستیم ۳) هنگامی که حجم داده‌های آموزشی زیاد باشد به علت محدود بودن حافظه، قادر به بار کردن آن در حافظه نخواهیم بود[۱].

۱-۲- بیان مساله تحقیق

در یادگیری ماشین، داده‌های آموزشی نقش مهمی را در تعیین کارایی الگوریتم یادگیری ایفا می‌کنند. تعداد این داده‌های آموزشی در طول زمان افزایش می‌یابد وداده‌های جدیدی از راه می‌رسد. این داده‌های جدید، ممکن است دید جدیدی از داده‌های قبلی را ارائه دهند یا توزیع آماری داده‌ها را تغییر دهند. از این رو، فهمیدن اهمیت داده‌های جدید و اجازه به این داده‌ها، برای نقش داشتن در آموزش، به منظور افزایش کارایی سیستم یادگیر، کاری بس زیرکانه است. آموزش دوباره‌ی سیستم یادگیر با همه‌ی داده‌ها، و کنار گذاشتن همه‌ی تجربه‌ی یادگیری گذشته، یکی از راه‌حل‌ها برای این مسئله است ولی این روش دارای نقص‌هایی می‌باشد ازجمله اینکه این

روش نا کارآمد است و همچنین، این روش قادر به نگه‌داری دانش قبلی نمی‌باشد. از این رو با توجه به این موضوع، ما بر آن شدیم که سیستمی را ارائه نماییم کـه قـادر باشـد دانـش یـاد گرفتـه شده‌ی قبلی را حفظ نماید در عـین حـال، قـادر بـه یـادگیری از داده‌هـای جدیـد نیـز باشـد. چنـین سیستمی، علاوه بر این مزیت‌ها باید بتواند از دانش یاد گرفته شده‌ی قبلی در دسته‌بندی متون فعلی استفاده کند و علاوه بر آن باید بتواند دقت بالایی نیز در دسته‌بندی متون داشته باشد.

۱-۳- اهداف و رویکردها

با توجه به مطالب ذکر شده در مورد اهمیت یادگیری افزایشی، هدف ما در این پایان‌نامه ارائـه سیستمی بودکه قادر باشد علاوه بر حفظ دانش قبلی داده‌های جدید را نیز فرا گیرد از این رو چنین سیستمی بایستی دارای قابلیت‌های زیر باشد:

- چنین سیستمی بایستی قابلیت بسط و گسترش داشته باشد؛ بدین معنی کـه قابلیـت بـالایی در یادآوری داده‌های یاد گرفته شده‌ی قبلی داشته باشد در عین حال بتواند در هنگام یـادگیری داده‌های جدید، داده‌هایی از داده‌های یاد گرفته شده‌ی قبلـی را بـه نحـوی در ایـن آمـوزش دخالت دهد که خط تصمیم ساخته شده در هر مرحلـه، بـه خـط بهینـه‌ی سراسـری نزدیـک شود.
- چنین سیستمی باید قادر باشد نقاط ضعف دانش یاد گرفته‌ی شده‌ی قبلی را بهبـود بخشـد و این نقاط ضعف را پوشش دهد.
- چنین سیستمی باید بتواند قابلیت خود تطبیقی داشته باشد بدین معنـی کـه در حـین آمـوزش فرضیه‌های ضعیف را شناسایی کرده و آنها را با فرضیه‌های قویتر جایگزین نماید.
- این سیستم باید بتواند بعد از اتمام مراحل آموزشی و در مرحله‌ی تسـت نهـایی فرضـیه‌هـای ایجاد کننده‌ی نویز را شناسایی نماید.
- این سیستم باید در مرحله‌ی تست نهایی با توجه به اینکه همه‌ی فضای داده‌ای را در اختیار دارد باید قادر باشد تا جایی که امکان دارد کارایی سیستم را افزایش دهد.
- دارای کارایی بالایی نسبت به روش‌های ارائه شده‌ی قبلی باشد.

۱-۴- دستاوردهای پایان‌نامه

در راستای رسیدن به اهداف پایان نامه چهار روش بصورت زیر ارائه شد:

در روش اول، ما سیستم یادگیری افزایشی ارائه نمودیم که در این سیستم، برای هر دسته‌ی آموزشی از یک دسته‌بند ماشین بردار پشتیبان برای یادگیری آن استفاده نمودیم. سپس، برای دسته‌ی آموزشی بعدی علاوه بر دسته‌ی آموزشی آن دسته، از بردارهای پشتیبان دسته‌های آموزشی قبلی به اضافه‌ی داده‌های غلط دسته‌بندی شده‌ی مجموعه‌های تصدیق دسته‌های آموزشی قبلی استفاده نمودیم. سپس، این دسته‌بندهای آموزش دیده در هر مرحله را ذخیره نمودیم. در پایان به منظور ارزیابی کارایی سیستم پیشنهادی یک مجموعه‌ی تست نهایی را به هر کدام از دسته‌بندهای آموزش دیده ارائه نمودیم و برای تولید خروجی نهایی، از روش اکثریت آراء ساده استفاده نمودیم. استفاده از بردارهای پشتیبان دسته‌های آموزشی قبلی، به منظور استفاده از تجربیات گذشته‌ی دسته‌بندهای پیشین در دسته‌بندهای کنونی و ساخت خط تصمیم بهینه بود و همچنین، استفاده از داده‌های غلط دسته‌بندی شده، برطرف کردن نقاط ضعف دسته‌بندهای پیشین وساخت دسته‌بندهای قوی بود.

در روش دوم برای اینکه بتوانیم میزان معیارهای ارزیابی کارایی را افزایش و خطای دسته‌بندی را کاهش دهیم به جای استفاده از تنها یک دسته‌بند برای هر دسته‌ی آموزشی، از چند دسته‌بند برای هر دسته‌ی آموزشی استفاده نمودیم. همچنین، به جای در نظر گرفتن وزن یکسان به هر دسته‌بند برای تاثیر در ترکیب خروجی‌ها، از روش اول پایان‌نامه‌ی آقای علی دانش استفاده نمودیم که ایشان به ازای ترکیب هر دسته و دسته‌بند از وزن مستقلی استفاده نموده‌اند و به منظور بروزرسانی مداوم این وزن‌ها در هر مرحله از یادگیری افزایشی، مجموعه تصدیق فعلی به اضافه‌ی تمام مجموعه تصدیق‌های قبلی را به همه‌ی دسته‌بندهای فعلی و همه‌ی دسته‌بندهای قبلی ارائه نمودیم که این امر سبب ثبات نسبی این وزن‌ها گردید.

در روش سوم، به منظور کامل کردن روش قبلی سیستم یادگیری افزایشی ارائه نمودیم که قادر بود برای هر دسته‌ی آموزشی از میان مجموعه‌ی دسته‌بندها، چندین دسته‌بند قابل قبول را انتخاب نماید. هدف از ارائه این روش کنارگذاشتن دسته‌بندهای ضعیف و جایگزین کردن آنها با دسته‌بندهای قویی که با داده های آموزشی بیشتری آموزش دیده بودند بود که این امر اشاره به قابلیت خود تطبیقی سیستم دارد.

در روش چهارم با بسط روش سوم در هنگام گرفتن خروجی نهایی، نودهایی را حذف و نودهایی را اضافه نمودیم. هدف از ارائه این روش حذف نودهای ایجاد کننده‌ی نویز که باعث پایین آمدن کارایی سیستم می‌شد بود و اضافه کردن نودهایی بود که دارای کارایی بالایی بودندکه این امر سبب افزایش کارایی کلی سیستم یادگیر تا جایی که امکان داشت می‌شد.

۱-۵- ساختار پایان نامه

این پایان نامه در شش فصل تنظیم گردیده است که رئوس مطالب و موضوعات اصلی هر فصل در ذیل بطور خلاصه بیان گردیده است:

- **فصل اول** به انگیزه، اهمیت و شرح موضوع تحقیق می‌پردازد. همچنین ساختار کلی پایان نامه و خلاصه مطالب ارائه شده در این تحقیق نیز بیان می‌گردد.

- **فصل دوم** در این فصل ابتدا به مطالعه‌ی نحوه بازنمایی اسناد بصورت بردارهای عددی می‌پردازیم سپس، به مطالعه‌ی انواع روش‌های انتخاب ویژگی فیلتر می‌پردازیم در ادامه، به بررسی دسته‌بندهای ارائه شده در این پایان‌نامه می‌پردازیم و در آخر به مرور کارهای ارائه شده‌ی قبلی در این مورد می‌پردازیم.

- **فصل سوم** ما در این فصل به ارائه‌ی دو روش می پردازیم. در روش اول ما سیستم یادگیری افزایشی ارائه نمودیم که در این سیستم، برای هر دسته‌ی آموزشی از یک دسته‌بند ماشین بردار پشتیبان برای یادگیری آن استفاده نمودیم و سپس، برای دسته‌ی آموزشی بعدی علاوه بر دسته‌ی آموزشی آن دسته، از بردارهای پشتیبان دسته‌های آموزشی قبلی به اضافه‌ی داده‌های غلط دسته بندی شده‌ی مجموعه‌های تصدیق دسته‌های آموزشی قبلی استفاده نمودیم. سپس، این دسته‌بندهای آموزش دیده در هر مرحله را ذخیره نمودیم و در پایان به منظور ارزیابی کارایی سیستم پیشنهادی یک مجموعه‌ی تست نهایی را به هر کدام از دسته‌بندهای آموزش دیده ارائه نمودیم و برای تولید خروجی نهایی، از روش اکثریت آراء ساده استفاده نمودیم. در روش دوم برای اینکه بتوانیم میزان معیارهای ارزیابی کارایی را افزایش و خطای دسته‌بندی را کاهش دهیم به جای استفاده از تنها یک دسته‌بند برای هر دسته‌ی آموزشی، از چند دسته‌بند برای هر دسته‌ی آموزشی استفاده نمودیم و همچنین، به

جای در نظر گرفتن وزن یکسان به هر دسته‌بند برای تاثیر در ترکیب خروجی‌ها، از روش اول پایان‌نامه‌ی آقای علی دانش استفاده نمودیم که ایشان به ازای ترکیب هر دسته و دسته‌بند از وزن مستقلی استفاده نموده‌اند و به منظور بروزرسانی مداوم این وزن‌ها در هر مرحله از یادگیری افزایشی، مجموعه تصدیق فعلی به اضافه‌ی تمام مجموعه تصدیق‌های قبلی را به همه‌ی دسته‌بندهای فعلی و همه‌ی دسته‌بندهای قبلی ارائه نمودیم

- **فصل چهارم** در این فصل ما سعی در بهبود دو روش ارائه شده در فصل قبل پرداختیم و بدین ترتیب دو روش جدید مطرح نمودیم. در روش سوم، به منظور کامل کردن روش قبلی، سیستم یادگیری افزایشی ارائه نمودیم که قادر بود برای هر دسته‌ی آموزشی از میان مجموعه‌ی دسته‌بندها، چندین دسته‌بند قابل قبول را انتخاب نماید. در روش چهارم با بسط روش سوم در هنگام گرفتن خروجی نهایی، نودهایی را حذف و نودهایی را اضافه نمودیم.

- **فصل پنجم** این فصل از پنج قسمت اصلی تشکیل شده است دربخش اول به معرفی مجموعه‌های داده‌ای می‌پردازیم سپس، در بخش دوم به معرفی معیارهای ارزیابی پرداخته و در بخش سوم و چهارم به ترتیب به مقایسه‌ی روش‌های پیشنهادی با یکدیگر و به مقایسه‌ی روش‌های پیشنهادی با روش‌های قبلی می‌پردازیم و سپس، در بخش آخر به تحلیل وزن دسته‌بندها که براساس روش اول پایان‌نامه‌ی آقای علی دانش وزن‌دهی شده‌اند می‌پردازیم.

- **فصل ششم** در این فصل راهکارهای ارائه شده در این پایان‌نامه بطور خلاصه بررسی شده و در ادامه پیشنهاداتی برای توسعه و بهینه‌سازی هر چه بیشتر این سیستم ارائه می‌شود.