دانشکده فنی

پایان نامه کارشناسی ارشد

رشته مهندسی صنایع- صنایع

عنوان پایان نامه :

دسته بندی داده ها به روش ماشین های بردار پشتیبان با توابع هدف چندگانه

نگارش : علی ندائی

استاد راهنما : دکتر علی محمد احمدوند

تابستان ۹۱


بسم الله الرحمن الرحيم

صورت جلسه دفاع

تقدیم به

پدر و مادر عزیزم

تشکر و قدردانی:

از خدای بزرگ برای سلامتی و پشتکاری که به من عنایت فرمود تا توانایی پشت سر گذاشتن فراز و فرودها و به سرانجام رساندن بخشی از اهداف زندگیم را بدست آورم سپاسگزارم.

از پدر و مادر عزیزم و سایر اعضای خانواده ام کمال تشکر را دارم و از محضر خدای متعال برایشان آرزوی توفیق روز افزون می نمایم. این اثر ناچیز را به پاس سال ها تلاش پدر و مادرم به محضرشان تقدیم می کنم.

همچنین از جناب آقای دکتر احمدوند سپاسگذارم که با راهنمایی ها و کمک های بی دریغشان، نه تنها در این پایان نامه و مقطع تحصیلی و نه فقط از جهت علمی، بلکه برای تمام طول عمرم درس های فراوانی به من آموختند. از خدای متعال برای ایشان و خانواده شان موفقیت و شادکامی روز افزون خواستارم.

چکیده:

ماشین بردار پشتیبان روشی بر پایه مفاهیم تئوری یادگیری آماری می باشد که در آن با استفاده از یک مدل برنامه ریزی ریاضی ابرصفحه ای برای دسته بندی داده ها یافته می شود. در این پژوهش مدلی برای ماشین بردار پشتیبان با نرخ خطای چندگانه ارائه می گردد که در آن خطای دسته بندی در دسته اول و دوم با هم برابر نمی باشند. مدل مذکور در حالت خاص معادل ماشین بردار پشتیبان معمولی بوده و هرگونه تفاوت نرخ خطا، بین دو مدل ایجاد اختلاف خواهد نمود.

پس از معرفی و مدلسازی، روش مورد بررسی با استفاده از مجموعه داده هایی تحت آزمایش قرار گرفته و نتایج نهایی از لحاظ دقت دسته بندی و هزینه های ناشی از خطای دسته بندی گزارش خواهند گردید.

واژگان کلیدی- ماشین بردار پشتیبان- دسته بندی- نرخ خطای چندگانه- توابع کِرنِل- ماشین بردار پشتیبان با حساسیت هزینه

فهرست مطالب

فهرست اشکال

فهرست جداول

فصل اول

کلیات تحقیق

۱-۱ مقدمه

یادگیری ماشینی[۱] یکی از حوزه های میان رشته ای است که هدف اصلی آن یافتن روش هـای حـل و ابـداع الگـوریتم هایی است که در حل مسائل طبیعی و روزمره راه گشا می باشند. ماشین بردار پشتیبان[۲] از زیر شـاخه هـای یـادگیری ماشینی بوده و بر اساس مفاهیم تئوری یادگیری آماری[۳] پایه گذاری شده اسـت. بیشـترین محبوبیـت ماشـین بـردار پشتیبان در تشخیص دستخط[۴] و تشخیص صدا[۵] می باشد که از حیث دقت در این حوزه از قویترین شبکه های عصبی نیز پیشی می گیرد. مدل اولیه ماشین بردار پشتیبان توسط وپنیک[۶] (۱۹۹۸) مطرح شد. مدل مذکور یک مـدل برنامـه ریزی درجه دو[۷] می باشد که در شرایط دو دسته ای به کار می رود. هر چند که این روش در سـال هـای اخیـر مـورد توجه بسیاری از پژوهشگران بوده است، اما علیرغم این حقیقت، همچنان خلاهایی در حوزه مذکور به چشم می خـورد.

[1] Machine learning
[2] Support Vector Machine (SVM)
[3] Statistical learning theory
[4] Handwriting recognition
[5] Voice recognition
[6] Vapnik
[7] Quadratic programming

لذا مدل ماشین بردار پشتیبان در حالت های توسعه یافته تر معمولا با توابع هدف چندگانه مدل می شود. این اصطلاح (ماشین بردار پشتیبان با توابع هدف چندگانه) در ادبیات موضوع مرسوم بوده و به حالت هایی اطلاق می شود که مدل مسئله علاوه بر هدف اصلی، اهداف دیگری را نیز دنبال می کند و انگیزه استفاده از آن نزدیک تر نمودن مدل به دنیای واقعی می باشد. در این فصل در مورد کلیات این تحقیق مطالبی ذکر خواهد شد. همچنین در مورد جنبه های نوآوری طرح مباحثی مطرح می گردید.

۱-۲ تعریف مسئله و بیان موضوعات تحقیق

دسته بندی، به دنبال یافتن مدلی پیشگویانه[1] می باشد که با استفاده از آن مدل، بتوان در مورد مسائل سازمان در سطوح مختلف تصمیم گیری نمود. این ابزار با توجه به کاربرد قابل ذکر، همواره به عنوان یکی از اصلی ترین تکنیک های موثر در تصمیم گیری در نظر گرفته می شود. روش های زیادی برای دسته بندی معرفی شده اند که مهمترین و پرکاربردترین آن ها در یک مقاله علمی- مروری توسط کاتسیانتیس[2] (۲۰۰۷) مورد بررسی قرار گرفته است. به طور خلاصه می توان به درخت تصمیم، شبکه عصبی، الگوریتم های بیزی و ... اشاره نمود.

یکی از نوین ترین و دقیق ترین روش های دسته بندی، ماشین بردار پشتیبان می باشد. در واقع این روش در ابتدا برای دسته بندی داده ها معرفی گردید و مدتی بعد از آن برای کشف الگوهای موجود در میان داده ها برای حالت های توسعه یافته تر مورد استفاده قرار گرفت (وپنیک، ۱۹۹۸). هدف اصلی در روش ماشین های بردار پشتیبان، حداقل نمودن خطای دسته بندی می باشد که این کار با حداکثر نمودن حاشیه دسته بندی صورت می پذیرد. بنابراین آنچه که در مدل ماشین بردار پشتیبان به عنوان تابع هدف ملاحظه می شود، حداکثر نمودن حاشیه است نه حداکثر نمودن دقت دسته بندی.

هدف از انجام این تحقیق یافتن پاسخ سوالات زیر است:

۱ -تابع هدف مرسوم در روش ماشین های بردار پشتیبان، نرخ خطای دسته بندی را یکسان در نظر می گیرد. این مطلب ممکن است برخی از واقعیت های موجود را خدشه دار کند. به عنوان مثال یک بانک را در نظر بگیرید که در

[1] Predictive Model
[2] Kotsiantis

آن می توان مشتریان متقاضی وام را به سه دسته پرخطر، کم خطر و بی خطر تقسیم بندی نمود. اگر یک شخص بی خطر را به اشتباه در دسته پرخطر قرار داده و به شخص مذکور وام تعلق نگیرد ممکن است هزینه تحمیل شده کمتر از حالتی باشد که به یک شخص پرخطر به اشتباه وام تعلق گیرد. بنابراین آیا می توان مدلی با در نظر گرفتن نرخ خطای چندگانه ارائه نمود؟

۲ -رویکردهای متفاوتی برای در نظر گرفتن نرخ خطای چندگانه در الگوریتم های دسته بندی معرفی شده اند. از آن جمله می توان به مقاله ارائه شده توسط لی[1] و همکاران (۲۰۰۹) اشاره نمود. آیا امکان بهبود چنین رویکردهایی برای ماشین بردار پشتیبان وجود دارد؟

۳ -در مدل اصلی ماشین بردار پشتیبان، متغیرهای مصنوعی موجود در قیود و تابع هدف به گونه ای تعریف شده اند که اگر مقدار آنها بزرگتر از ۱ باشد، نقطه متناظر دارای خطا در دسته بندی می باشد. رویکردهای مورد بحث تاکنون، نقاطی که متغیر مصنوعی متناظر آنها دارای مقداری بین صفر و یک می باشد را نیز در نظر می گیرند. در حالیکه این نقاط دچار خطای دسته بندی نمی باشند و بنابراین نباید در تعیین محل ابرصفحه دسته بند موثر واقع شوند.

۱-۳ مفروضات تحقیق

۱ - فرآیند پیش پردازش[2] بر روی داده ها انجام شده است. بنابراین داده ی مخدوش، ناقص، ناسازگار و... وجود ندارد.

۲ - مشخصه های[3] مورد استفاده برای داده ها در دسته بندی موثر می باشند. بنابراین نیاز به حذف هیچکدام از مشخصه ها (بدلیل عدم اهمیت آن ها) نمی باشد.

۳ - مقادیر موجود در کلیه مشخصه ها برای همه ی داده ها به صورت قطعی اند. بنابراین حالت احتمالی[4]، تصادفی[5]، فازی[6] و هر حالتی که فرض قطعیت را مختل نماید در میان داده ها وجود ندارد.

۴ - خطای اندازه گیری و خطای انسانی در میان داده ها وجود ندارد.

---

[1] Li
[2] Pre-processing
[3] Features
[4] Probabilistic
[5] Stochastic
[6] Fuzzy

۵ - مشخصه ها از هم مستقل اند.

۶ - هزینه ی جمع آوری داده ها (پر نمودن مشخصه ها[1]) در نظر گرفته نمی شود (برابر صفر می باشد).

۷ - داده های مورد استفاده از سایت های مرتبط جمع آوری شده اند. این داده ها به صورت متعادل[2] می باشند. یعنی تقریبا نیمی از آن ها متعلق به دسته ۱ و نیمی دیگر متعلق به دسته ۲ می باشند. بنابراین نیازی به متعادل سازی داده ها نخواهد بود.

۱-۴ ضرورت انجام تحقیق

به طور کلی از روش های پژوهش عملیاتی به منظور افزایش دقت و ایجاد یک روش قانون محور[3] در دسته بندی استفاده می شود. از برنامه ریزی ریاضی می توان برای ایجاد مدل هایی با محدودیت های خاص تر استفاده نمود. به عنوان مثال کاریزوسا[4] و باراگان[5] (۲۰۰۶) و مالدونادو[6] و همکاران (۲۰۱۱) با استفاده از تکنیک برنامه ریزی ریاضی محدودیت هایی را به مدل های دسته بندی اضافه نموده اند. هدف از افزودن چنین محدودیت هایی بالا بردن دقت و یا جنبه واقع گرایانه مدل می باشد. به طور کلی دو نوع تقابل بین روش های دسته بندی و روش برنامه ریزی ریاضی وجود دارد:

۱ - روش هایی که بر اساس مفاهیم بهینه سازی و پژوهش عملیاتی توسعه داده شده اند و بدون چنین مفاهیمی بی معنی هستند. مثل ماشین های بردار پشتیبان. در چنین حالت هایی مفاهیم پژوهش عملیاتی از روش مورد استفاده جدا پذیر نیستند.

۲ - روش هایی که مفاهیم بهینه سازی و پژوهش عملیاتی در آن ها به عنوان یک روش جدید به کار رفته است. مثلا درخت تصمیم که از روش های پژوهش عملیاتی نیز ساخته می شود.

---

[1] Features
[2] Balance
[3] Rule-based
[4] Carrizosa
[5] Baragan
[6] Maldonado

با توجه به فعالیت های انجام شده در حوزه ماشین بردار پشتیبان به نظر می رسد با ارائه مدل های برنامه ریزی ریاضی ذکر شده در قسمت ۱-۲ می توان قابلیت های مدل ماشین بردار پشتیبان را به عنوان یکی از پرکاربردترین روش های دسته بندی به واقعیت نزدیک تر نمود. روش های توضیح داده شده می توانند کارایی دسته بندی و صحت آن را بهبود بخشند.

۱ ۴ کاربردهای تحقیق

در این تحقیق مفاهیم اولیه در حوزه ماشین بردار پشتیبان به طور کامل توضیح داده شده است. بنابراین این تحقیق می تواند مرجع مناسبی برای مطالعه در حوزه ماشین بردار پشتیبان تلقی گردد.

از دیدگاه کاربردی نیز این تحقیق می تواند در کلیه مراکز و سازمان هایی که قصد طبقه بندی داده ها و استفاده از آن ها به منظور پیش بینی را دارند نیز مورد استفاده قرار بگیرد.

۱ ۵ روش های جمع آوری اطلاعات

روش های جمع آوری اطلاعات شامل جستجو در سایت های اینترنتی معتبر علمی، مطالعه مقالات و کتاب ها و مشاوره با استاد راهنما و استفاده از تجربیات و پژوهش های ایشان است. همچنین مشاوره با اساتید این حوزه با نظارت و مشورت استاد راهنما کمک زیادی در گردآوری اطلاعات نموده است. داده های مربوط به شبیه سازی نیز از پایگاه اطلاعاتی مربوط به ماشین بردار پشتیبان به آدرس http://archive.ics.uci.edu/ml/ استخراج گردیده است. داده های مذکور با فرمت *txt*. دریافت شده و سپس به فرم ماتریسی به نرم افزار *MATLAB* منتقل شده و با فرمت *mat*. مورد تحلیل قرار گرفته اند. داده های مذکور در مقالات متعدد و معتبر به کار گرفته شده است و دسترسی به مقالاتی که از هر یک از داده ها استفاده نموده اند امکان پذیر می باشد. به علاوه به منظور تسهیل بررسی نتایج محاسبتی این تحقیق، نام هر یک از مجموعه داده های[1] استفاده شده نیز ذکر شده است.

[1] Datasets

۱ ۶ نحوه شبیه سازی

شبیه سازی و نتایج محاسباتی موجود در این تحقیق با استفاده از نرم افزار *MATLAB R2010a* و الگوریتم های موجود در جعبه ابزار آن انجام شده است. این الگوریتم ها عبارتند از الگوریتم نقطه درونی و روش بهینه سازی متوالی درجه دو که با استفاده از آن مسئله برنامه ریزی درجه دو با محدودیت تبدیل به یک مسئله بدون محدودیت می شود. همچنین پس از حل مسئله با الگوریتم های مطروحه، در فاز دوم با استفاده از یک الگوریتم ابتکاری به بهبود جواب پرداخته شده است. در برخی موارد برای اطمینان از صحت عملکرد نرم افزار، خروجی به دست آمده با خروجی حاصل از سایر منابع و نرم افزارها مقایسه شده است.

۱ ۷ ساختار پایان نامه

در فصل دوم پایان نامه، در ابتدا به تشریح مفاهیم اولیه مورد نیاز مدل ماشین بردار پشتیبان پرداخته شده است. هدف از این قسمت آشنایی خواننده با مدل و نحوه عملکرد ماشین بردار پشتیبان می باشد. مرور ادبیات و پژوهش های مرتبط با این تحقیق در ادامه فصل ۲ معرفی گردیده و مدل هایی که تاکنون ارائه شده اند مورد بحث و بررسی قرار خواهند گرفت.

در فصل سوم نرخ خطای چندگانه[۱] به عنوان یک رویکرد جدید در ماشین بردار پشتیبان مورد بررسی قرار خواهد گرفت. به عبارتی برای این حالت برخلاف حالت های مرسوم، نرخ خطا به صورت یک ماتریس مربعی خواهد بود، بطوریکه مرتبه این ماتریس برابر تعداد دسته های موجود و عناصر قطر اصلی آن همگی برابر صفر می باشند. برای حل این مسئله در فاز اول از الگوریتم های حل مسئله برنامه ریزی درجه دوم (مثل شرط بهینگی کاروش کوهن- تاکر[۲] یا روش لاگرانژین[۳] و یا روش های نقطه درونی[۴] و الگوریتم برنامه ریزی درجه دو متوالی[۵]) استفاده شده و در فاز دوم نیز از یک الگوریتم ابتکاری[۶] جواب به دست آمده بهبود داده می شود.

[1] Multiple error rate
[2] Karush Kohn-Tucker Condition
[3] Lagrange
[4] Interior point
[5] Sequential quadratic programming
[6] Heuristic

در فصل چهارم نیز به حل مدل های مذکور و مقایسه آن ها با روش های مرسوم پرداخته شده است. همچنین در مورد روش های حل هر یک از مدل ها توضیحاتی ارائه می شود. مرتبه[1] زمانی الگوریتم ها نیز مورد تفسیر قرار خواهند گرفت. در انتهای فصل چهارم با تغییر در نرخ خطای دسته بندی و انجام آنالیز حساسیت[2]، مقدار تغییرات در دقت دسته بندی و خطای ایجاد شده در هر دسته مورد بررسی شده و مباحث اصلی خاتمه خواهند پذیرفت.

نهایتا فصل پنجم از این تحقیق به جمع بندی مباحث و پیشنهاداتی برای مطالعات آتی خواهد پرداخت.

۱ ۸ نوآوری های پایان نامه

کلیه مدل های ارائه شده در فصل سوم به همراه نتایج محاسباتی مربوطه در فصل چهارم از جنبه های نوآوری این تحقیق محسوب می شود. تا کنون مدلی برای نرخ خطای چندگانه با در نظر گرفتن نقاطی که دارای خطای دسته بندی می باشند ارائه نشده است. بنابراین مدل ارائه شده در فصل سوم و الگوریتم ابتکاری معرفی شده برای بهبود جواب بدست آمده از جنبه های نوآوری طرح می باشد.

[1] Order
[2] Sensitivity analysis