مقایسه روش‌های نزدیکترین همسایه مجاور، SVM، C۴.۵ و نیو– بیز برای دسته‌بندی داده‌ها

مهدی نصیری[۱]، نوید کاردان[۲]، علی هادیان[۳]، بهروز مینایی[۴]

چکیده

در این مقاله دقت روش‌های نزدیک‌ترین همسایه مجاور(k-NN)، C۴.۵، SVM و NB برای دسته‌بندی داده‌ها با توجه به معیارهای "دقت" و "سطح زیر منحنی" مقایسه شده است. تاثیر عواملی مانند اندازه مجموعه داده، تعداد صفات دودویی و عددی و درصد داده استفاده شده برای آزمون در شرایط استقلال مجموعه داده از مسئله خاص، مورد بررسی قرار گرفته است.

نتایج بدست آمده بیانگر آن است که k-NN و SVM در اکثر موارد بهتر از ۲ روش دیگر عمل می‌نماید. برای یک داده خاص هر چه تعداد داده استفاده شده برای داده آموزش بیشتر باشد، تمام روش‌ها در اکثر موارد نتیجه بهتری را ارائه می‌دهد. این تاثیر در روش SVM بیشتر است.

کلمات کلیدی

روش نیو- بیز، نزدیکترین همسایه مجاور، SVM، C۴.۵، مجموعه داده‌ها، داده‌های آموزشی، داده‌های آزمایشی

۱. مقدمه

انتساب اشیاء به دسته‌های مربوطه را دسته‌بندی[۱] می‌نامند. هدف از دسته‌بندی یافتن مدلی برای صفتِ دسته، به‌عنوان تابعی از سایر متغیرهاست، تا بتوان بوسیله آن، نوع دسته را برای رکود داده‌های قبلا دیده نشده و با بیشترین دقت ممکن تعیین کرد [۱]. روش‌های گوناگونی برای دسته‌بندی وجود دارد که از آن جمله می‌توان به درخت‌های تصمیم‌گیری (DT)[۲] ، روش‌های مبتنی بر قانون[۳]، رگرسیون لاجستیک (LR) ، Naïve Bayes (NB) ، Support Vector Machine (SVM) ، نزدیکترین همسایه مجاور (k-NN) ، شبکه های عصبی و ... اشاره کرد [۱۲,۱۱,۱۰,۹,۳,۱].

هر کدام از روش‌های فوق دارای مزایا و معایبی می‌باشند و بسته به ویژگی‌های "مجموعه داده"ها دقت[۴] و صحت[۵] متفاوتی از خود بروز می‌دهند. منحنی ROC معیاری متداول برای تشخیص توان و قدرت[۶] روش‌های مختلف دسته‌بندی است و از سطح زیر منحنی (AUC) برای مقایسه روش‌ها استفاده می‌شود [۵,۴]. از معیارهای دیگری نیز مانند G-mean و دقت در [۸,۶] و از معیار کارآیی RMSE در [۹] استفاده شده است. مقایسه روش-های دسته‌بندی از جمله موضوعات مورد بحث در [۹,۷] می باشد.

در این مقاله با استفاده از داده‌ها استاندارد موجود ۴ روش k-NN، NB، SVM، C۴.۵ را با توجه به معیارهای دقت و سطح زیر منحنی مقایسه می‌نماییم. در بخش دوم، به معرفی ۴ روش دسته‌بندی می‌پردازیم. مقایسه روش‌های دسته‌بندی و نتایج حاصل از این مقایسه موضوع بخش سوم می-باشد. در بخش چهارم نتیجه‌گیری و کارهای آینده، و در انتها مراجع آمده است.

[۱] دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه علم و صنعت ایران nasiri@comp.iust.ac.ir
[۲] دانشجوی کارشناسی ارشد هوش مصنوعی دانشگاه علم و صنعت ایران n_kardan@comp.iust.ac.ir
[۳] دانشجوی کارشناسی مهندسی نرم‌افزار دانشگاه علم و صنعت ایران hadian@comp.iust.ac.ir
[۴] استادیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت ایران b_minaei@iust.ac.ir

۲. معرفی ۴ روش دسته بندی

۲.۱. روش NB

استفاده از تئوری بیز، ابزاری قدرتمند برای تصمیم‌گیری در شرایط عدم‌قطعیت[۷] است. یک شکل خیلی ساده از دسته‌بندِ بیز تحت عنوانِ نیو- بیز نامیده می‌شود که به صورت زیر عمل می‌نماید [۲].

اگر D را مجموعه رکودها در نظر بگیریم که هر رکورد دارای برداری شامل n صفت باشد، هدف یافتنِ مقداری برای صفت دسته است که مقدارِ عبارت (۱) را حداکثر کند.

$$P(C_i|A_1, A_2, \dots, A_n) \qquad (۱)$$

که این امر، هم‌ارز با حداکثر شدن فرمول (۲) است:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad = P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_n|C_i) \qquad (۲)$$

در اینجا p احتمال پسین[۸] می‌باشد.

۲.۲. روش k-NN

اکثر دسته‌بندهای داده شامل دو بخش اساسی می باشند:

مرحله استقراء[۹] برای ساختن مدل دسته‌بندی از داده‌ها

مرحله استنتاج[۱۰] برای اِعمال مدل ساخته شده بر روی نمونه‌های جدید و از قبل دیده نشده.

در این نوع دسته‌بندها، مدل بلافاصله بعد از دیدن مجموعه آموزش ساخته می‌شود و سپس مدلِ حاصل، روی مجموعه آزمایش اعمال می‌گردد. این رهیافت‌ها "یادگیرنده‌های‌مشتاق[۱۱]" نامیده می‌شوند. استراتژی دیگری نیز وجود دارد که عمل عمومی‌سازی[۱۲] از داده آموزش را تا زمان استفاده از آن به می‌باشد که داده آموزش Rote classifierتعویق می‌اندازد. به این روش‌ها، "یادگیرنده‌های‌تنبل[۱۳]" گفته می‌شود. یک نمونه از این نوع دسته‌بندها، ورودی را در حافظه خود ذخیره می‌کند و تعیین دسته را صرفاً زمانی انجام می‌دهد که داده آزمایش ورودی مشابه با یکی (یا چند نمونه) از داده‌های آموزش ذخیره شده باشد، بنابراین داده آزمایش می‌تواند بر اساس نزدیکترین همسایه خود دسته‌بندی و تعیین دسته شود. این روش، ایده اصلی در بُعدی برخورد می‌کند. در نتیجه می‌توان n با هر کدام از داده‌ها به‌عنوان یک نقطه در فضای NN) می‌باشد. NNدسته‌بندِ نزدیکترین همسایه مجاور[۱۴] (فاصله نقاط از همدیگر را معیار نزدیکی آنها قرار داده و مشابه‌ترین نقاط را انتخاب نمود. برای تعیین فاصله، می‌توان از فاصله اقلیدسی استفاده کرد:

$$d(p,q) = \sqrt{\sum_{i=1}^{d}(p_i - q_i)^2} \qquad (۳)$$

بنابراین، k همسایه مجاور (k-NN) برای یک نقطه z از مجموعه آزمایش، عبارت خواهد بود از k نقطه از مجموعه آموزش که کمترین فاصله را تا نقطه z دارند [۳].

۲.۳. روش SVM

این روش از جمله روش‌های نسبتاً جدیدی است که در سال‌های اخیر کارایی خوبی نسبت به روش‌های قدیمی‌تر برای دسته‌بندی از جمله شبکه‌های عصبی پرسپترون نشان داده است. مبنای کاری دسته‌بندی کنندهٔ SVM دسته‌بندی خطی داده‌ها است و در تقسیم خطی داده‌ها سعی می‌کنیم خطی را انتخاب کنیم که حاشیه اطمینان بیشتری داشته باشد. حل معادلهٔ پیدا کردن خط بهینه برای داده‌ها به وسیله روش‌های QP که روش‌های شناخته شده‌ای در حل مسائل محدودیت‌دار هستند صورت می‌گیرد. قبل از تقسیمِ خطی برای اینکه ماشینِ ما بتواند داده‌های با پیچیدگی بالا را دسته‌بندی کند داده‌ها را به وسیله‌ی تابعِ phi به فضایی با ابعاد خیلی بالاتر می‌بریم. برای اینکه بتوانیم مساله ابعاد خیلی بالا را با استفاده از این روش‌ها حل کنیم از قضیه دوگانی لاگرانژ برای تبدیلِ مساله‌ی مینیمم‌سازی مورد نظر به فرم دوگانی آن که در آن به جای تابع پیچیده‌ی phi که ما را به فضایی با ابعاد بالا می‌برد، تابعِ ساده‌تری به نامِ تابع هسته که ضرب برداری تابع phi است ظاهر می‌شود استفاده می‌کنیم. از توابع هسته مختلفی از جمله هسته‌های نمایی، چندجمله‌ای و سیگموید می‌توان استفاده نمود.

SVM اصولاً یک ماشین خطی است که ایدهٔ اصلی آن ایجاد یک فوق صفحه به عنوان سطح تصمیم گیری می باشد، به طوری که حد تفکیک بین نمونه های مثبت و منفی حداکثر شود . این روش با استفاده از یک شیوه که بر پایه تئوری آموزش آماری بنا نهاده شده، به خصیصه های بهینه فوق دست پیدا می‌کند. به صورت دقیقتر SVM یک پیاده سازی تقریبی از روش "حداقل کردن ریسک ساختاری" است.

ساختار الگوریتم آموزش SVM بر اساس یک هسته ضرب داخلی بین یک بردار پشتیبان مثل xi و بردار x به دست آمده از فضای ورودی است. کوچک‌ترین زیر مجموعه داده‌های آموزش که توسط الگوریتم فوق استخراج شده‌اند، بردار تقویت[15] نامیده می‌شوند. بسته به اینکه این هسته ، ضرب داخلی چگونه تولید شود، ممکن است مشی نهایی آموزش مختلفی با صفحات تصمیم گیری غیرخطی مربوطه بدست آیند. نمونه آموزش $\{(x_i, d_i)\}_{i=1}^{N}$ را در نظر بگیرید که در آن xi یک الگوی ورودی برای نمونه iام و di پاسخ مطلوب (خروجی نهایی) متناظر با ورودی فوق است. فرض می‌کنیم که الگوهای (دسته‌های) نمایش داده شده با زیر مجموعه۱ − = di و ۱ + = di تفکیک‌پذیر خطی باشند . معادله یک صفحه مجموعه تصمیم‌گیری به فرم فوق صفحه به صورت زیر است:

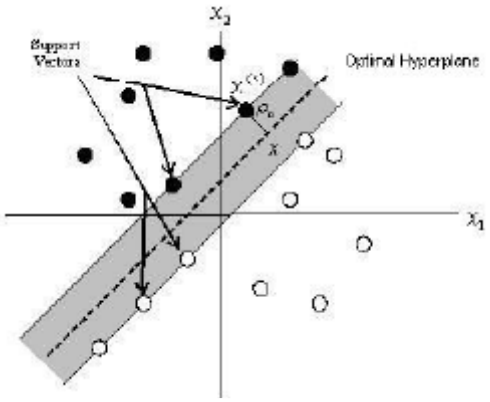
$w^T x + b = ۰$

که در آن x یک بردار ورودی ، w یک بردار وزن قابل تنظیم و b یک بایاس است. معادله فوق را می‌توان به صورت زیر نوشت:

$W^t x_i + b >= ۰$ for $d_i = +۱$

$W^t x_i + b < ۰$ for $d_i = -۱$

برای یک بردار وزن w و یک بایاس b معین، فاصله بین فوق صفحه تعریف شده در معادله (۴) و نزدیک‌ترین نقطه داده‌ای، حد تفکیک نامیده شده و با ρ نشان داده می‌شود. هدف SVM پیدا کردن فوق صفحه منحصر به فردی است که حد تفکیک در آن ماکزیمم شود. در این وضعیت سطح تصمیم‌گیری به عنوان فوق صفحه بهینه در نظر گرفته می‌شود . شکل زیر ساختار هندسی یک فوق صفحه بهینه را برای یک فضای ورودی دو بعدی نشان می‌دهد.



شکل۱.ساختار هندسی یک فوق صفحه بهینه

حال b_o و w_o را به ترتیب مقادیر بردار وزن و بایاس بهینه در نظر بگیرید. بنابراین فوق صفحه بهینه به صورت زیر تعریف می‌گردد:

$W_o^t x + b_o = ۰$

بحث ا صلی پیدا کردن پارامترهای b_o و w_o برای فوق صفحه بهینه با مجموعه

آموزش معین $\tau = \{(x_i, d_i)\}$ باید در شرط زیر(xo ,do) صدق کنند:

$W_o^t x_i + b_o >= ۰$ for $d_i = +۱$

$W_o^t x_i + b_o < ۰$ for $d_i = -۱$

است. پس زوج نقاط داده ویژه (xi , di) که در شرط تساوی هردوسطرفرمول بالا صدق می‌کنند، Support Vectorنامیده شده و ازینرو به آنها Support Vector Machineگفته می شود. این بردارها نقش برجسته ای در عملکرد این نوع از ماشینهای آموزش بازی می کنند. به صورت مفهومی بردارهای پشتیبان نقاط داده ای هستند که در نزدیکی صفحه تصمیم گیری واقع شده اند و بنابراین به سختی دسته بندی می شوند. برای به دست آوردن فوق صفحه بهینه، با کمک ضرایب لاگرانژ محاسباتی انجام می‌شود و تابعی که باید بیشینه شود.

۴.۲. روش C۴.۵

C۴.۵ یک معیار استاندارد در یادگیری ماشین است. انتخاب صفت در ID۳ و C۴.۵ بر اساس حداقل کردن مقیاس اطلاعات در یک گره است. هر مسیر از ریشه به سمت یک گره، نمایان‌گر یک قانون دسته‌بندی می‌باشد.

تئوری بر این اساس است که تعداد آزمون‌هایی که باعث می‌شود یک نمونه جدید در داخل پایگاه داده، دسته‌بندی شود، حداقل گردد. بخش انتخاب صفت در C٤.٥ بر این اساس است که پیچیدگی درخت تصمیم به شدت وابسته به مقدار اطلاعاتی است که با آن صفت در ارتباطند. با انتخاب آن صفت، اطلاعات بیشتر از هر صفت دیگری، جدا و تقسیم می‌شوند. الگوریتم C٤.٥ دامنه دسته‌بندی را علاوه بر صفات قیاسی در انواع صفات عددی نیز توسعه می‌دهد. الگوریتم اصولا صفتی را که حداکثر درجه جداسازی بین دسته‌ها را دارد را انتخاب می‌کند و درخت تصمیم را بر اساس آن می‌سازد.

تولید درخت تصمیم اولیه از روی مجموعه داده‌ای، مهم‌ترین بخش الگوریتم C٤.٥ می‌باشد. الگوریتم در نهایت یک دسته‌بند را در قالب یک درخت تصمیم تولید می‌کند که دارای ۲ نوع گره است. یک گره بصورت برگ که یک دسته را مشخص می‌کند و یک گره تصمیم که آزمون‌هایی روی یک صفت انجام می‌دهد تا یک شاخه یا زیر درخت به ازای هر خروجی آزمون تولید می‌کند.

روش ساخت درخت مشابه‌ای، بصورت بازگشتی به هر زیر مجموعه از نمونه‌ها اعمال می‌شود. این رویه ادامه می‌یابد تا زیر مجموعه‌ها شامل نمونه-هایی باشند که به یک دسته تعلق داشته باشند.

فرآیند ساخت درخت، یک فرآیند واحد نمی‌باشد. متاسفانه، مشکل پیدا کردن کوچکترین درخت تصمیم از روی یک نمونه داده‌ای مساله‌ای NP-Complete است. بنابراین، باید روش‌های ساخت درخت غیر عقب‌گرد[۱۶] باشند و بصورت حریصانه عمل نمایند.

۳. مقایسه روش‌های دسته‌بندی و نتایج حاصل

روش‌های دسته‌بندی k-NN، SVM، NB و C۴.۵ روی ۴ مجموعه داده اعمال شد، به‌صورتی که برای محاسبه دقت هر یک از مجموعه داده‌ها، ۴ بار با ۹۰٪، ۷۰٪، ۳۰٪ و ۱۰٪ از داده برای داده آموزشی و بقیه برای آزمایش استفاده شد. در جدول ۱ نیز ویژگی‌های داده‌ها ارائه شده است.

پس از محاسبه معیارهای دقت و سطح زیر منحنی برای آنها و میانگین‌گیری، نتایج مطابق جدول‌های ۳و ۲ بدست آمد.

در جدول ۳، ۷۰٪ داده‌ها را برای داده‌های آموزشی و ۳۰٪ را برای داده‌های آزمایشی استفاده کرده‌ایم.

در روش k-NN، از مقدار ۳ برای متغییر k و فاصله اقلیدسی برای محاسبه نزدیک‌ترین استفاده کرده‌ایم. همچنین در روش NB برای تخمین احتمال شرطی از Relative Frequency استفاده شده است. در روش SVM نیز از هسته RBF(نمایی) استفاده شده است.

با توجه به نتایج حاصل و نمودارها می‌توان گفت:

- برای مجموعه داده‌های با رکودهای کم، مقادیر بدست آمده برای هر کدام از روش‌های دسته‌بندی دارای تغییرات زیادی می-باشند. ولی با افزایش تعداد رکودها، حالت پایدارتری حاصل می‌گردد.
- روش NB برای حالتی که نسبت متغیرهای اسمی به عددی زیاد است (حالات اول و دوم) دارای دقت برابر روش K-NN می-باشد.
- روش k-NN در ۱۵ موقع بهتر از NB عمل می‌نماید.
- روش K-NN برای حالتی که تعداد متغیرهای اسمی زیاد می‌شود دارای دقت پایینی است.
- در حالتی که تمام متغیرها عددی است دقت روش‌های K-nn و SVM به خصوص با مجموعه داده زیاد با هم اختلاف زیادی ندارند بطوری‌که در ۹۰٪ و بیشتر بعنوان داده آزمون با هم برابر هستند.
- روش C٤.٥ برای حالتی که تعداد صفات دودویی بیشتر است دارای سطح زیر نمودار بالایی است.
- تمام روش‌ها برای حالتی که نوع صفات فقط عددی است دارای سطح زیر نمودار کمتری است. این کاهش در C٤.٥ بیشتر است.
- تمام روش‌ها برای حالتی که نوع صفات فقط عددی است دارای دقت کمتری است.
- به جز نوع داده‌های دسته چهارم در بقیه نوع داده‌ها روش NB دارای سطح زیر نمودار تقریبا برابری است.
- روش C۴.۵ برای حالتی که نوع صفات فقط عددی است دارای کمترین سطح زیر نمودار است.
- روش K-NN برای حالتی که تعداد صفات اسمی بیشتر است دارای کمترین سطح زیر نمودار است.
- روش C۴.۵ برای حالتی که صفات فقط عددی است دارای کمترین سطح زیر نمودار است.

- روش NB برای حالتی که تعداد صفات دودویی بیشتر است دارای دقت کمتری است.
- هر دو روش میزان اعتبار یک روش تقریبا به یک نسبت تغییر می‌کند.
- در حالتی که میزان استفاده داده برای آموزش ۷۰٪ است نسبت تغییرات سطح زیر نمودار و دقت دقیق‌تر از ۳ حالت دیگر است.
- با توجه به دقت می‌توان K-NN و SVM را بهتر از ۲ روش دیگر نام برد.
- با توجه به سطح زیر نمودار می‌توان SVM را بهتر از ۳ روش دیگر نام برد.

در جدول۳، ۷۰٪ داده‌ها را برای داده‌های آموزشی و ۳۰٪ را برای داده‌های آزمایشی استفاده کرده‌ایم:

جدول ۱- ویژگی داده‌ها

تعداد دسته	تعداد صفت اسمی[۱۷]	تعداد صفت دودویی	تعداد صفت عددی	تعداد فرا صفت	تعداد داده	مجموعه داده
۳	۰	۰	۴	۰	۱۵۰	اول
۴	۴	۰	۲	۰	۱۷۲۸	دوم
۷	۰	۱۶	۱	۱	۱۰۱	سوم
۶	۰	۰	۹	۰	۲۱۴	چهارم

جدول ۲- دقت پیش‌بینی دسته‌بندها

NB	C۴.۵	SVM	k-NN	مجموعه داده
۹۲.۶۷	۹۲.۶۷	۹۵.۳۳	۹۵.۳۳	اول
۹۳.۱۱	۹۴.۲۳	۹۶.۹۰	۹۵.۱۱	
۹۱.۸۱	۹۳.۷۱	۹۵.۷۲	۹۴.۴۸	
۹۰.۷۴	۸۸.۳۰	۸۷.۵۷	۹۱.۹	
۸۵.۶۶	۹۲.۶۱	۹۶.۷۶	۸۳.۱۸	دوم
۸۴.۹۷	۹۰.۳۹	۹۶.۴۴	۸۳.۸۲	
۸۴.۰۳	۸۵.۲۳	۹۳.۱۵	۸۳.۷۱	
۸۱.۶۲	۷۸.۸۸	۸۷.۳۹	۷۸.۹۰	
۹۲.۷۴	۹۵.۴۶	۹۴.۵۶	۹۲.۷۳	سوم
۸۸.۷۱	۹۳.۲۳	۹۰.۰۰	۹۴.۸۴	
۸۲.۲۹	۸۹.۴۳	۶۸.۰۰	۹۶.۰۰	
۶۶.۴۲	۷۴.۳۴	۴۲.۸۹	۸۹.۴۴	
۶۹.۹۰	۶۸.۶۴	۷۰.۹۱	۶۸.۱۸	چهارم
۶۸.۷۶	۶۴.۶۹	۶۸.۱۳	۶۸.۲۹	
۶۵.۱۳	۶۰.۵۳	۵۵.۵۴	۶۴.۰۰	
۵۶.۴۱	۵۱.۶۲	۴۵.۸۳	۵۷.۹۳	

جدول۳ – مقدار سطح زیر منحنی

NB	C۴.۵	SVM	k-NN	مجموعه داده
۹۸.۸۱	۹۷.۰۵	۹۹.۹۷	۹۸.۲۶	اول
۹۷.۶۹	۹۷.۷۳	۹۹.۷۱	۹۱.۵۱	دوم
۹۷.۵۵	۹۹.۱۷	۹۹.۵۴	۹۹.۶۱	سوم
۹۱.۶۳	۸۲.۰۶	۸۹.۸۶	۸۸.۳۴	چهارم

۴. نتیجه‌گیری

با توجه به نتایج بدست آمده و مقایسه دقت و سطح زیر منحنی روش‌های فوق می‌توان گفت که در اکثر موارد روش‌های k-NN و SVM عملکرد بهتری دارند.

نتایج به‌دست آمده در این مقاله بر اساس داده‌های استاندارد مورد استفاده در جهان بوده و نیازمند درستی‌یابی به‌وسیله داده‌هایی با حجم‌های بیشتر و واقعی‌تر از لحاظ هیچ‌مقدار بودن و ... می‌باشد تا صحت آنها مورد تأیید قرار گیرد. در هر حال نتایج به‌دست آمده برای مطالعه ما- به‌عنوان اولین مطالعه مقایسه‌ای بین ۴ روش فوق و بر اساس نوع صفات- معنی‌دار می‌باشند. از جمله زمینه‌های موجود برای کارهای آینده می‌توان به مقایسه دسته‌بندهای دیگر بخصوص دسته‌بندهای پویا بر روی مجموعه داده‌های با تعداد صفات بیشتر، وجود داده با مقادیر نامعلوم و استفاده از معیارهای دیگر به‌جای دقت و سطح زیر منحنی ، اشاره کرد.

۵. مراجع

[۱] Ian H. Witten, Eibe Frank; "*Data Mining: Practical Machine Learning Tools and Techniques*", Second Edition; Elsevier, Morgan Kaufmann publications; ۲۰۰۵.

[۲] J. Han, M. Kamber; "*Data Mining: Concepts and Techniques*", Second edition, Elsevier, ۲۰۰۶.

[۳] Pang-Ning Tan, Michael Steinbach, Vipin Kumar; "*Introduction to Data Mining*", Addison-Wesley, ۲۰۰۶.

[۴] Adam Fadlalla; "*An experimental investigation of the impact of aggregation on the performance of data mining with logistic regression*", Elsevier, Information & Management, pp. ۶۹۵–۷۰۷, ۲۰۰۵.

[۵] J. Huang, J. Lu,C.X. Ling; "*Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy*", Proceedings of the Third IEEE International Conference on Data Mining , ۲۰۰۳.

[۶] L. Xu, M-Y. Chow, X. Z. Gao; "Comparisons *of Logistic Regression and Artificial Neural Network on Power Distribution Systems Fault Cause Identification*", IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications, Finland, June ۲۸-۳۰, ۲۰۰۵.

[۷] N.B. Amor, S. Benferhat, Z. Elouedi; "*Naive Bayes vs decision trees in intrusion detection systems*", Proceedings of the ۲۰۰۴ ACM Symposium on Applied Computing, Nicosia, Cyprus, ۲۰۰۴.

[۸] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura; "*A comparative study of K-Nearest Neighbour, Support Vector Machine and Multi-Layer Perceptron for Thalassemia screening*", Chemometrics and Intelligent Laboratory Systems, pp. ۱۳– ۲۰, ۲۰۰۳.

[۹] Yong Soo Kim; "*Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size*", Elsevier, Expert Systems with Applications, pp.۱۲۲۷–۱۲۳۴, ۲۰۰۸.

[۱۰] S. Haykin, "*Neural Networks: A Comprehensive Foundation*", Second Edition, Prentice-Hall Inc., ۱۹۹۹.

[۱۱] C. J. C. Burger, "*A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery*", Bell Labs/Lucent, ۲(۲):۹۵۵-۹۷۴, ۱۹۹۸.

[۱۲] V. N. Vapnick, "*The Nature of Statistical Learning Theory*", Second Edition, Springer-Verlag New York Inc., ۲۰۰۰.

---

[۱] Classification

[۲] Decision Tree

[۳] Rule Based Methods

[۴] Precision

[۵] Accuracy

[۶] Robustness

[۷] uncertainty

[٨] posterior probability

[٩] inductive

[١٠] deductive

[١١] Eager learners

[١٢] Generalization

[١٣] Lazy learners

[١٤] Nearest Neighbor

[١٥] Support Vector

[١٦] non-backtracking

[١٧] Nominal