Iranian Journal of Medical Informatics (2012) Vol 2, Issue 2

Assessment of Classification Algorithms in the Diagnosis of Diabetes and Breast Cancer

ارزیابی الگوریتم های دسته بندی در تشخیص دیابت و سرطان سینه

Seyed Abbas Mahmoodi, Maryam Sadat Mahmoodi, Vahide Sadat Abbasniya, Seyed Mostafa Mahmoodi

Abstract — Nowadays medical sciences and physicians are faced with the volume of data. Since the diagnosis is not always easy, therefore the physician should consider results of the patient tests and decisions taken in the past for patients with similar conditions in order to make a good decision. In other words, the physician will need knowledge and experience. However, due to the large number of patients and any patient's multiple tests, the need for an automated tool to explore the former patients is felt. One of the important methods used to derive data is data mining . The aim of this paper is the application of classification algorithms for the diagnosis of diabetes and breast cancer and identifying the best algorithm. By comparing the obtained results, it turns out that there is no algorithm with maximum efficiency[1].

Keywords — data mining, classification, diabetes, breast cancer.

۱. چکیده

امروزه علوم پزشکی و پزشکان با حجم زیاد داده ها روبه رو می باشند. ازآنجایی که تشخیص بیماری همواره کار آسانی نیست بنابراین پزشک برای اتخاذ یک تصمیم مناسب، باید نتیجه‌ی آزمایش‌های بیمار و تصمیم‌هایی که در گذشته برای بیماران با وضعیت مشابه گرفته است، را بررسی کند. به عبارت دیگر پزشک نیازمند دانش و تجربه خواهد بود. ولی به دلیل تعداد زیاد بیماران و آزمایش‌های متعدد هر بیمار، نیاز به یک ابزار خودکار برای کاوش در میان بیماران قبلی احساس می‌شود. یکی از این روش‌های مهم که برای استنتاج داده‌ها استفاده می‌شود، داده‌کاوی است. هدف این مقاله این است که الگوریتم های دسته بندی را برای تشخیص درست بیماری دیابت و سرطان سینه بکاربرد و بهترین الگوریتم را شناسایی نماید. که با مقایسه نتایج به دست آمده معلوم می شود که هیچ الگوریتمی وجود ندارد که همواره دارای کارایی بیشینه باشد.

کلمات کلیدی: داده‌کاوی، دسته‌بندی، دیابت، سرطان سینه.

۲. مقدمه

بیماری دیابت یکی از خطرناک ترین بیماری ها در قرن حاضر است ،به طوری که قاتل خاموش نامیده می شود. این بیماری یک تهدید اساس برای سلامت جوامع در کشورهای پیشرفته ودر حال توسعه به شمار می آید ودر حال حاضر دیابت چهارمین علت مرگ ومیر در بیشتر کشورهای توسعه یافته می باشد .دیابت یک بیماری مزمن است که وقتی بدن توان تولید انسولین مورد نیاز سلول ها راندارد یاهنگامی که بدن نمی تواند از انسولین تولید شده ،استفاده موثر داشته باشد ،اتفاق می افتد .توقف تولید انسولین یا استفاده نکردن از انسولین هر دو باعث افزایش گلوکز در خون می شود. دیابت دارای دو نوع اصلی است: نوع اول و نوع دوم .افرادی که دیابت نوع اول دارند، بدنشان مقدارکمی انسولین تولید می کنند ویا اصلا انسولینی در بدنشان تولید نمی شود و لازم است که برای ادامه زندگی انسولین به بدنشان تزریق کنند. این نوع دیابت بیشتر در کودکان و نوجوانان رخ می دهد. نوع دوم بیماران دیابتی، افرادی هستند که بدنشان توان استفاده موثر از انسولین مورد نیاز بدن را ندارند. این نوع دیابت در افراد بالای ۳۰سال رخ می‌دهد و عوارض بسیار خطرناک‌تری نسبت به نوع اول دارد. دیابت نوع دوم یا دیابت ملیتوس نسبت به نوع اول بسیار شایع‌تر است [1] .میلیون‌ها نفردر سرتاسر جهان به این بیماری مبتلا هستند و متاسفانه افراد بسیار بیشتری از ابتلای خود به این بیماری و یا احتمال بالای ابتلاء به این بیماری اطلاع ندارند و ممکن است سال‌ها طول بکشد تا از ابتلای خود به دیابت اطلاع پیدا کنند. این در حالی است که زمان، برای پیشگیری و درمان دیابت بسیار حیاتی است. از مهمترین عوارض دیابت می‌توان به حمله‌های قلبی، نابینایی، بیماری‌های کلیوی، فشارخون و آسیب‌های عصبی اشاره کرد.

سرطان سینه رایج ترین شکل سرطان در زنان می‌باشد. علل سرطان سینه و یا زمینه سازان این سرطان عبارتند از: عوامل ژنتیکی و نژادی، برنامه غذایی و چاقی، هورمونها، تابش اشعه، سرطانهای شیمیایی، سن بالا، موقعیت جغرافیایی. تشخیص دیابت و سرطان سینه و یا آگاهی یافتن از احتمال بالای ابتلا به این بیماری‌ها همواره

[1] S.A. Mahmoodi is with Department of Computer Engineering, Islamic Azad University, Science and Research Branch, Yazd, Iran (email: sa_mahmoodi_85@yahoo.com)

M.S. Mahmoodi is with Department of Computer, Islamic Azad University, Ferdows Branch, Ferdows, Iran (email: m_mahmoodi_64@yahoo.com)

V.S. Abbasniya is with Department of Physiology, Islamic Azad University, Ferdows Branch, Ferdows, Iran (email: abbasnia.vahideh@yahoo.com)

S.M. Mahmoodi is an Assistant Professor, Department of Oral Pathology, Faculty of Dentistry, Birjand University of Medical Sciences, Birjand, Iran (email: mahmoudi_m16@yahoo.com).

کار آسانی نیست. چرا که این بیماری‌ها علائم متعددی را بروز می‌دهند که بعضی از این علائم در سایر بیماری‌ها نیز وجود دارد. بنابراین پزشک برای اتخاذ یک تصمیم مناسب، باید نتیجه‌ی آزمایش‌های بیمار و تصمیم‌هایی که در گذشته برای بیماران با وضعیت مشابه گرفته است، را بررسی کند. به عبارت دیگر پزشک نیازمند دانش و تجربه خواهد بود. ولی به دلیل تعداد زیاد بیماران و آزمایش‌های متعدد هر بیمار، نیاز به یک ابزار خودکار برای کاوش در میان بیماران قبلی احساس می‌شود[2].یکی از این روش‌های مهم که برای استنتاج داده‌ها استفاده می‌شود، داده‌کاوی می‌باشد.

در این مقاله قصد داریم با پیاده سازی چند الگوریتم توسط نرم افزار Weka و مقایسه الگوریتم ها توسط معیارهای اندازه گیری روی دو مجموعه داده‌ی دیابت و سرطان سینه الگوریتم بهینه را تعیین نماییم.

۳. توصیف دو مجموعه داده

در این بخش به توصیف دو مجموعه داده‌ی دیابت و سرطان سینه که از مخزن UCI، که مرجعی برای یادگیری ماشین می‌باشد، خواهیم پرداخت[3].

أ. مجموعه داده ی دیابت

مجموعه داده Pima شامل ۹ ویژگی و ۲ کلاس می‌باشد. مجموعه کلاس‌ها به دو گروه خوش‌خیم و بدخیم تقسیم می‌شود، که کلاس اول شامل ۵۰۰ نفر افراد سالم و کلاس دوم شامل ۲۶۸ نفر بیمار می‌باشند. جدول ۱ صفات مربوط به این مجموعه داده همراه با توصیف مختصری از آن‌ها را نشان می دهد.

جدول ۱: ویژگی های مجموعه داده Pima

نام ویژگی	مینیمم	ماکسیمم	میانگین
Number Pregnant	۰	۱۷	۳/۸۴
Plasma	۰	۱۹۹	۱۲۰/۸۹
Blood Pressure	۰	۱۲۹	۶۹/۱۱
Skin fold	۰	۹۹	۲۰/۵۶
Serum Insulin	۰	۸۴۶	۷۹/۷۴
Body mass Index	۰	۶۷	۳۲
Pedigree Function	۰/۷۸.	۲/۴۲	۰/۴۶
Age	۲۱	۸۱	۳۳

ب.مجموعه داده سرطان سینه

مجموعه داده Wisconsinشامل ۹ ویژگی و ۲ کلاس می‌باشد. مجموعه کلاس ها به دو گروه خوش‌خیم و بدخیم تقسیم می‌شود، که کلاس اول شامل ۴۴۱ نفر افراد سالم و کلاس دوم شامل ۲۴۰ نفر بیمار می‌باشند. جدول ۲ صفات مربوط به این مجموعه داده همراه با توصیف مختصری از آن‌ها را نشان می دهد.

جدول ۲: ویژگی های مجموعه داده Wisconsin

نام ویژگی	مینیمم	ماکسیمم	میانگین
Radius	۱	۱۰	۴/۴۵
Texture	۱	۱۰	۳/۲
Perimeter	۱	۱۰	۳/۲۴
Area	۱	۱۰	۲/۹
Smoothness	۱	۱۰	۳/۲
Compactness	۱	۱۰	۳/۶
Concave	۱	۱۰	۳/۴
Symmetry	۱	۱۰	۲/۹
Fractal Dimension	۱	۱۰	۱/۶

به منظور مقایسه روش پیشنهادی با سایر روش‌ها، در اینجا از روش‌های1-NN، K-NN ، C4.5 ، RBF Network و NaiveBayes استفاده شده است.

۴. الگوریتم های دسته بندی

به منظور مقایسه روش پیشنهادی با سایر روش‌ها، در اینجا از روش‌های1-NN، K-NN ، C4.5 ، RBF Network و NaiveBayes استفاده شده است.

أ. مجموعه داده ی دیابت

الگوریتم KNN برای دسته‌بندی نمونه‌ها استفاده می‌شود، و براساس این اصل است که یک نمونه با K نمونه که خصوصیات مشابه بیشتری با هم دارند، دسته‌بندی می‌شود. به این صورت که K تا از نمونه‌های نزدیک را برای نمونه جدید شناسایی می کند و برچسب کلاسی که بیشترین تکرار را در میان این نمونه‌ها دارد به عنوان کلاس نتیجه‌ی آن مشخص می‌شود. بنابراین باید معیاری را برای تعیین فاصله بین نمونه‌ها مشخص کرد. این فاصله، باید فاصله‌ی بین نمونه‌های یک کلاس را مینیمم کند و فاصله‌ی بین نمونه‌های کلاس‌های متفاوت را ماکزیمم نماید.

ب. الگوریتم C4.5

از مهم‌ترین الگوریتم‌های درخت تصمیم می‌توان به الگوریتم قدرتمند C4.5 اشاره کرد. این الگوریتم از یک معیار مبتنی بر آنتروپی استفاده می‌نماید. همچنین از تکنیک‌های هرس کردن برای از بین بردن شاخه های اضافی استفاده می‌کند.

ج. الگوریتم RBF Network

یکی از روش‌های رایج دسته‌بندی مبتنی بر پرسپترون، شبکه عصبی و از آن جمله شبکه عصبی RBF می‌باشد. شبکه‌های عصبی RBF از سه لایه تشکیل می-شوند، لایه ورودی، لایه نهان که نرون‌های آن دارای تابع RBF می‌باشند و یک لایه

خروجی.

د. الگوریتم NaiveBayes

یکی از مهم‌ترین روش‌های دسته‌بندی آماری می‌باشد. شبکه های NB، شبکه های بیزی خیلی ساده‌ای هستند که از گراف‌های بدون دور جهت‌دار، با تنها یک والد و چندین فرزند تشکیل شده‌اند و فرض می‌کند که نودهای فرزندان مستقل هستند. این الگوریتم احتمال شرطی هر ویژگی داده شده را با توجه به دسته مربوطه‌اش یاد می‌گیرد. سپس عمل دسته‌بندی با بکاربردن قوانین بیز برای محاسبه مقدار احتمالی دسته نتیجه نمونه داده شده، با دقت بالایی انجام می‌شود[۴]

۵. نرم افزار مورد استفاده برای مقایسه

نرم‌افزارهای فراوانی برای داده‌کاوی و یادگیری ماشین در حوزه‌های مختلف داده ها موجود می‌باشند. هریک از آن‌ها با توجه به نوع اصلی داده‌هایی که مورد کاوش قرار می‌دهند، روی الگوریتم‌های خاصی متمرکز شده‌اند.

میزکار Weka ، مجموعه‌ای از الگوریتم‌های روز یادگیری ماشین و ابزارهای پیش‌پردازش داده‌ها می‌باشد. این نرم‌افزار به گونه‌ای طراحی شده است که می‌توان به سرعت، روش‌های موجود را به صورت انعطاف‌پذیری روی مجموعه‌های جدید داده، آزمایش نمود.

نرم‌افزار Weka در دانشگاه Waikato واقع در نیوزلند توسعه یافته است و نام آن از عبارت "Waikato Environment for Knowledge Analysis" استخرج گشته است

۶. معرفی معیارهای ارزیابی الگوریتم پیشنهادی

چندین معیار برای ارزیابی کارایی الگوریتم پیشنهادی در این مقاله در نظر گرفته می‌شود که عبارتند از:

۱-زمان اجرا

۲-نرخ دسته‌بندی

۳- نرخ صحت

۴- نرخ فراخوانی

۵-F- سنجش

در ادامه به بررسی هر یک از معیارهای ذکر شده با توجه به جدول۳ می پردازیم.

جدول۳: ماتریس درهم ریختگی

نوع دسته		دسته تشخیص داده شده: مثبت	دسته تشخیص داده شده: منفی
دسته واقعی	مثبت	TP	FN
	منفی	FP	TN

دسته‌ها در مسأله‌ی تشخیص سرطان با دو دسته مثبت و منفی و چهار عدد TP، FP، FN و TN با توجه به نوع دسته مثبت و منفی محاسبه می‌گردند.

TP: شامل نمونه‌هایی است که توسط یک قانون پوشش داده می‌شوند و متعلق به کلاسی هستند که در قسمت مقدم آن قانون قرار دارد.

FP: شامل نمونه‌هایی است که توسط یک قانون پوشش داده می‌شوند و متعلق به کلاسی نیستند که در قسمت مقدم آن قانون قرار دارد.

FN: شامل نمونه‌هایی است که توسط یک قانون پوشش داده نشده‌اند و متعلق به کلاسی هستند که در قسمت مقدم آن قانون قرار دارد.

TN: شامل نمونه‌هایی است که توسط یک قانون پوشش داده نشده‌اند و متعلق به کلاسی نیستند که در قسمت مقدم آن قانون قرار دارد[5].

نرخ دسته‌بندی بیانگر دقت الگوریتم پیاده‌سازی شده در دسته‌بندی دسته های مختلف موجود در مسأله‌ی تشخیص سرطان می‌باشد. این معیار در واقع درصد دسته‌بندی درست الگوریتم می‌باشد. به عبارتی نرخ دسته بندی، تعداد نمونه‌هایی می‌باشد که به درستی دسته‌بندی شده است، و از رابطه زیر بدست می‌آید:

$$Classificationrate = \frac{TP + TN}{TP + FP + FN + TN} \quad (۱)$$

نرخ صحت که برای هر کدام از دسته‌های موجود قابل محاسبه می‌باشد، جهت تعیین دقت دسته‌بندی برای هر کدام از دسته‌ها در نظر گرفته شده است. در واقع این معیار نشان دهنده‌ی درصد موفقیت روش دسته‌بندی کننده در تشخیص نمونه-های مربوط به هرکدام از دسته‌ها می‌باشد(رابطه۲). نرخ فراخوانی که همانند معیار قبل برای هر کدام از دسته‌های موجود محاسبه می‌گردد، درصد قابلیت اعتماد به خروجی روش دسته‌بندی کننده را نشان می‌دهد(رابطه ۳)[6,7].

به دلیل اینکه امکان بهبود هردو معیار ذکر شده، به طور همزمان کار مشکلی می-باشد، بایستی میان آن‌ها با رعایت کردن یک مصالحه، خروجی نهایی را ارزیابی نماییم. معیار F- سنجش، برای مصالحه میان دو معیار ذکر شده بکار می‌رود[8,9].

معیار F- سنجش، از رابطه ۴ بدست می‌آید.

$$precision = \frac{TP}{TP + FP} \quad (۲)$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \quad (۳)$$

$$F - Measure = \frac{2 * precision * \mathrm{Re}call}{precision + \mathrm{Re}call} \quad (۴)$$

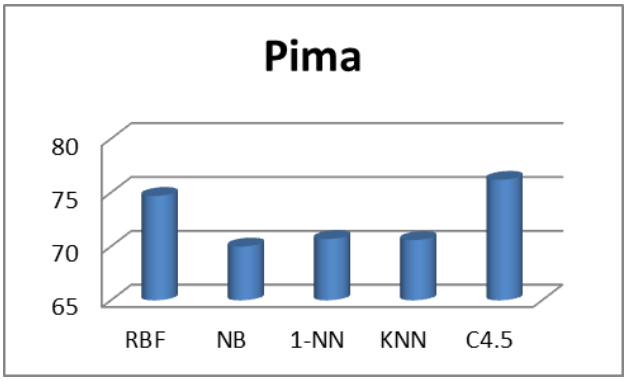
در این مقاله کلیه‌ی روش‌های دسته‌بندی‌کننده با توجه به معیارهایی که ذکر شده ارزیابی می‌گردند.

۷. نتایج

در این جا نتایجی را که توسط نرم افزار Weka برای دو مجموعه داده بیماری دیابت و سرطان سینه حاصل شده است را نشان می دهیم. جدول ۴ نتایج‌بدست-آمده‌برای‌مجموعه داده دیابت و شکل ۱ نمودار مقادیر دقت دسته بندی، و جدول ۵ نتایج‌بدست‌آمده‌برای‌مجموعه داده سرطان سینه و شکل ۲ نمودار مقادیر دقت دسته بندی ان را نشان می دهد. همان طور که دیده می شود الگوریتم RBF برای مجموعه داده دیابت و الگوریتم K-NN برای مجموعه داده سرطان سینه بهترین الگوریتم ها می باشند.

جدول۴: نتایج بدست آمده برای مجموعه داده Pima

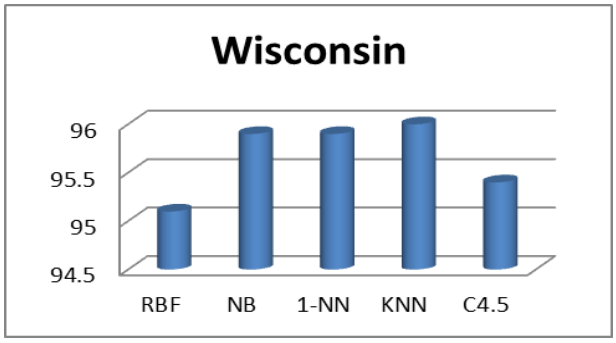
نام الگوریتم	دسته-نرخ بندی	صحت نرخ	نرخ فراخوانی	F- نرخ سنجش
RBF Network	٪۷۴/۷	٪۷۴/۷	٪۷۵/۷	٪۷۵
NaiveBayes	٪۷۰	٪۷۰/۵	٪۷۲/۵	٪۷۰/۸
1-NN	٪۷۰/۷	٪۷۰/۷	٪۷۰/۷	٪۷۰/۷
K-NN	٪۷۰/۶	٪۶۸/۴	٪۶۹/۶	٪۶۸/۴
C4.5	٪۷۵/۲	٪۷۵/۴	٪۷۵/۱	٪۷۵/۲



شکل۱: نمودار مقادیر پارامتر دقت دسته بندی الگوریتم ها در بیماری دیابت

جدول۵: نتایج بدست آمده برای مجموعه داده Wisconsin

نام الگوریتم	دسته-نرخ بندی	صحت نرخ	نرخ فراخوانی	F- نرخ سنجش
RBF Network	٪۹۵/۱	٪۹۵/۳	٪۹۵/۲	٪۹۵/۲
NaiveBayes	٪۹۵/۹	٪۹۵/۹	٪۹۵/۷	٪۹۵/۸
1-NN	٪۹۵/۹	٪۹۵/۹	٪۹۵/۹	٪۹۵/۹
K-NN	٪۹۶	٪۹۶/۱	٪۹۶	٪۹۶
C4.5	٪۹۵/۴	٪۹۵/۴	٪۹۵/۴	٪۹۵/۵



شکل۲: نمودار مقادیر پارامتر دقت دسته بندی الگوریتم ها در سرطان سینه

۸.نتیجه گیری

نتایج بدست آمده روی دو مجموعه داده بیماری دیابت و سرطان سینه نشان می دهد که در تشخیص بیماری ها هرگز نمی توان الگوریتمی را به عنوان الگوریتم بهینه معرفی نمود. در نتیجه برای هر کاربرد با توجه به مجموعه داده مورد استفاده می توان الگوریتمی را به عنوان الگوریتم بهینه معرفی نمود.

REFERENCES

1 Polat K, Gunes S, Arslan A. A cascade learning system for classification of diabetes disease: Genetalized Discriminate Analysis and Least Square Support Vector Machine. Expert Systems with Applications, 2010; 34: 482-487.

2 Andres C, Pena R, Sipper M. Designing breast cancer diagnostic systems via a hybrid Fuzzy-Genetic methodology. IEEE International Fuzzy Systems Conference, 2010; 1:135-139.

3 Howland J. Preventing Automobile Injury: New Findings From Evaluative Research. Dover, MA: Auburn House Publishing Company 1988:163–96.

4 Wu X, Kumar V, Quinlan JR. Top 10 Algorithms in Data Mining. Knowledge and Information Systems, 2009; 14: 1-37.

5 Ganji MF, Abadeh MS. Parallel Fuzzy Rule Learning Using an ACO-Based Algorithm for Medical Data Mining. IEEE Fifth International Conference on Bio-Innspired Compting: theories and Applications, 2010: 573-581.

6 Ganji MF, Abadeh MS. Using Fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease. Iranian conference of Electrical Engineering, ICEE, 2011.

7 Ganji MF, Abadeh MS. An Intelligence Fuzzy Classification System for Diabetes Detection. Iranian Conference on Fuzzy Systems, 2010.

8 Abadeh MS, Habibi J, Soroush E. Induction of Fuzzy Classification Systems via Evolutionary ACO-Based Algorithms. International Journal of Simulation, Systems, Science, Technology, 2008; 9:1-8.

9 Tsang CH, Kwong S, Wang H. Genetic-Fuzzy Rule Mining Approach and Evolution of Feature Selection Techniques for Anomaly Intrusion Detection. Pattern Recognition, 2009; 40: 2373-2391.