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

Automatic Detection of Erythemato-Squamous Diseases Using K-Nearest Neighbor Algorithm

تشخیص خودکار بیماری اریتماتو اسکاموز با استفاده از الگوریتم کای – نزدیکترین همسایه

Taha Samad Soltani, Mostafa Langarizadeh, Mostafa Shanbezadeh

Abstract — *Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases. One of the most important usage of data mining in machine learning domain is differential diagnosis. The differential diagnosis of erythemato-squamous diseases is a difficult problem in dermatology. The goal of this research is gaining to a high performance differential diagnosing of erythemato-squamous disease by using KNN Learning algorithms. UCI data base including 366 records has been used in this research. KNN algorithm was implemented and accuracy of system was compared with other methods used to diagnose such disease previously the highest accuracy was related to k=5. Sensitivity, specificity and accuracy in the best condition were respectively obtained 1, 1, and 0.98 and compared with other methods. KNN approach with preprocessing based on Relief-F strategy and Cross Fold modeling could be suggested as an effective data mining method to classify and diagnosis different diseases*[1].

Keywords — data mining, Machine Learning, K-nearest neighbor, erythemato-squamous diseases, differential diagnosis.

۱. چکیده

داده کاوی در مراقبت بهداشتی شاخه ی بسیار مهمی در تشخیص و فهم عمیق تر داده های پزشکی می باشد. داده کاوی بهداشتی در صدد حل مسائل دنیای واقعی در تشخیص و درمان بیماریها می باشد. یکی از مهمترین کاربردهای داده کاوی در حیطه یادگیری ماشین، تشخیص افتراقی می باشد و این نوع تشخیص در بیماری اریتماتواسکاموز یک چالش واقعی در درماتولوژی می باشد. هدف ازاین پژوهش تشخیص کارآمد این بیماری با استفاده از الگوریتم نزدیکترین همسایه که یکی از روشهای یادگیری ماشین است، می باشد. این پژوهش بر روی داده های استاندارد مجموعه داده *UCI* دانشگاه کالیفرنیا انجام شده که حاوی اطلاعات ۳۶۶ بیمار است استفاده شده است. الگوریتم نزدیکترین همسایه بر روی این داده ها پیاده سازی شد و میزان دقت و صحت سیستم در مقایسه با سایر روشهای انجام شده در تشخیص افتراقی بیماری اریتماتو اسکاموز محاسبه شد .از میان مقادیر مختلف همسایگی، بالاترین میزان صحت به ازای همسایگی ۵ بدست آمد. میزان حساسیت ، اختصاصی بودن و صحت در بهترین حالت به ترتیب برابر ۱،۱ و ۰،۹۸۰۱ بدست آمد. این روش با سایر پژوهش های انجام شده روی مجموعه داده استاندارد درماتولوژی مورد مقایسه قرار گرفت. رویکرد مبتنی بر الگوریتم کا نزدیکترین همسایه به همراه پیش پردازش بر مبنای استراتژی *Relief-F* و نمونه سازی *Cross Fold* به عنوان روشی کارآمد در هوش مصنوعی جهت داده کاوی با هدف طبقه بندی و تشخیص افتراقی بیماریها پیشنهاد می شود.

کلیدواژه ها: داده کاوی، یادگیری ماشین، کای نزدیکترین همسایه، بیماری اریتماتو اسکاموز، تشخیص افتراقی

[1] T.S. Soltani is a PhD student in Medical Informatics, Faculty of Allied Medical Sciences, Tehran University of Medical Sciences, Tehran, Iran (email: t-ssoltany@razi.tums.ac.ir)

M. Langarizadeh is with Tehran University of Medical Sciences, Tehran, Iran (corresponding author, email: mlangarizadeh@tums.ac.ir)

M. Shanbezadeh is a Master student in Health Information Technology, School of Health Management and Information Sciences, Tehran University of Medical Sciences, Tehran, Iran (email: messi9068@gmail.com).

۲. مقدمه

تشخیص افتراقی بیماری اریتماتواسکاموز یک چالش واقعی در درماتولوژی می باشد.[۱] با وجود تفاوت های بسیار اندک، تمام بیماریهای اریتماتو اسکاموزدر ویژگی های بالینی سرخی و اندازه مشترکند. ۶ گروه مختلف در این بیماری وجود دارند که عبارتند از پسوریازیس۲، سبوریک درماتیت۳، لیچن پلانوس۴، پیتیریاز روزا۵، درماتیت مزمن و روبرا پیلاریس۶ و این اشکال از بیماری غالبا در درمانگاههای پوستی دیده می شود. [۲] در تشخیص این بیماری بیوپسی بسیار مهم می باشد و متاسفانه این بیماری ویژگیهای هیستو پاتولوژی را نیز در بر می گیرد. [۳] مشکل دیگر در تشخیص افتراقی احتمال نمایش ویژگیهای بیماری دیگر در مرحله شروع بیماری و همچنین تداوم این نشانه ها در مراحل بعدی بیماری می باشد. در ابتدا بیماران بصورت بالینی از نظر ۱۲ ویژگی بررسی می شوند و سپس نمونه ی پوست اشخاص برای استخراج ۲۲ ویژگی هیستوپاتولوژی مورد ارزیابی قرار می گیرند. این ویژگیها با مطالعه ی پوست توسط میکروسکوپ انجام می پذیرد[۴]

داده کاوی گامی مهم در کشف و استخراج دانش می باشد این اصطلاح به معنای کاوش مجموعه داده های بزرگ برای استخراج الگوهای ناشناخته بین داده ها است.[۵] از آنجایی که کشف روابط بین داده ها در روش های سنتی آماری بسیار مشکل می باشد[۶]، کاربرد داده کاوی به سرعت در بخشهای وسیعی از قبیل سازمان های ارائه خدمات بهداشتی، پیش بینی مالی و پیش بینی هوا گسترش یافته است.[۷]

داده کاوی در مراقبت بهداشتی شاخه بسیار مهمی در تشخیص و فهم عمیق تر داده های پزشکی می باشد. داده کاوی بهداشتی در صدد حل مسائل دنیای واقعی در تشخیص و درمان بیماریها است. [۸] داده کاوی در مراقبت بهداشتی زمینه ی تحقیقاتی مهمی برای پیش بینی بیماریها و درکی عمیق تر از داده های بهداشتی محسوب می شود. داده کاوی بهداشتی قصد دارد تا مسائل بهداشتی دنیای واقعی در تشخیص و درمان بیماریها را حل نماید.[۹] پژوهشگران از این روش برای تشخیص بیماریهای مختلف استفاده می کنند و برای این کار از روشها و الگوریتم های مختلفی بهره می برند که میزان صحت و دقت آنها متفاوتی است [۱۰]

الگوریتم طبقه بندی کای- نزدیکترین همسایه روشی محبوب در تشخیص چندکلاسی در حیطه ی تشخیص الگو می باشد[۱۱] این روش یکی از پرکاربردترین روشهای داده کاوی و مسائل طبقه بندی است.[۱۲] این مقاله با بکارگیری الگوریتم کای- نزدیکترین همسایه در تشخیص افتراقی بیماری اریتماتواسکوماز، سعی در دستیابی به کارایی مناسبی جهت تشخیص افتراقی این بیماری دارد.

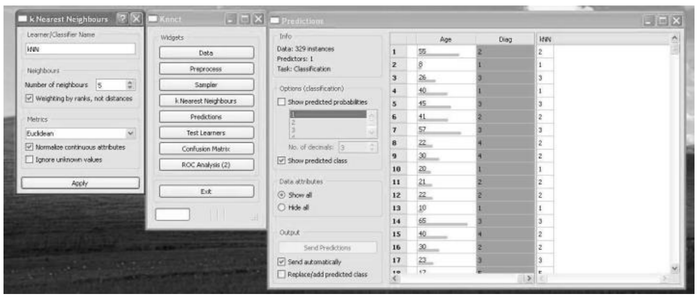
۳. کارهای انجام شده

در این پژوهش برای مدلسازی و انجام محاسبات مسئله، ۴ مرحله طی شد که در شکل ۱ نشان داده شده است و در ادامه تشریح می شوند. پیاده سازی و طراحی الگوریتم به زبان پایتون و در محیط کانواس Orange انجام پذیرفت که در شکل ۲ قسمتی از محیط برنامه و خروجی های مرتبط نمایش داده شده است.

[2] psoriasis
[3] seboreic dermatitis
[4] lichen planus
[5] pityriasis rosea
[6] rubra pilaris

أ. اکتساب داده

در این پژوهش از مجموعه داده درماتولوژی UCI دانشگاه کالیفرنیا استفاده شد.[۱۳] این مجموعه داده شامل ۳۶۶ رکورد و متشکل از ۳۴ ویژگی می باشد. برای هر بیمار ۱۲ ویژگی بالینی و ۲۲ ویژگی پاتولوژیکی ثبت شده است. این ویژگیها در جدول شماره ۱ نشان داده شده اند. هر یک از این ویژگیها دامنه مقادیر صحیح دارند. برای نمونه سابقه خانوادگی بیماری مقادیر ۰ یا ۱ را به عنوان عدم مشاهده یا مشاهده سابقه خانوادگی و فامیلی می گیرد. اکثر ویژگیها دامنه ای بین ۰ تا ۳ را دریافت می کنند که ۰ نشان دهنده عدم وجود و ۳ نشان دهنده بالاترین احتمال وجود آن عارضه یا ویژگی می باشد و۱ و۲ مقادیری بین آنها را می پذیرند. کلاسهای خروجی نیز شامل ۶ کلاس می باشند که نشان داده شده است. اطلاعات دیگری از قبیل تعداد نمونه ها در مجموعه داده نیز در جدول ۱ نمایش داده شده است.



شکل ۱ – محیط برنامه طراحی شده

شکل ۲ – مراحل انجام پژوهش

جدول ۱. مجموعه داده بیماری اریتماتو اسکاموز مورد استفاده در پژوهش حاضر

Erythemato-squamous diseases (number of patients)	Features	
	Clinical	Histopathological
Psoriasis (111)	Feature 1: Erythema	Feature 12: Melanin incontinence
Seboreic dermatitis (60)	Feature 2: Scaling	Feature 13: Eosrinophils in the infiltrate
Lichen planus (71)	Feature 3: Definite borders	Feature 14: PNL infiltrate
Pityriasis rosea (48)	Feature 4: Itching	Feature 15: Fibrosis of the papillary dermis
Chronic dermatitis (48)	Feature 5: Koebner phenomenon	Feature 16: Exocytosis
Pityriasis rubra pilaris (20)	Feature 6: Polygonal papules	Feature 17: Acanthosis
	Feature 7: Follicular papules	Feature 18: Hyperkeratosis
	Feature 8: Oral mucosal involvement	Feature 19: Parakeratosis
	Feature 9: Kneeand elbow involvement	Feature 20: Clubbing of the rete ridges
	Feature 10: Scalp involvement	Feature 21: Elongation of the rete ridges
	Feature 11: Family history	Feature 22: Thinning of the suprapapillary epidermis
	Feature 34: Age	Feature 23: Pongiform pustule
	Feature 34: Age	Feature 23: Pongiform pustule
		Feature 24: Munro microabcess
		Feature 25: Focal hypergranulosis
		Feature 26: Disappearance of the granular layer
		Feature 27: Vacuolization and damage of basal layer
		Feature 28: Spongiosis
		Feature 29: Saw-tooth appearance of retes
		Feature 30: Follicular horn plug
		Feature 31: Perifollicular parakeratosis
		Feature 32: Inflammatory mononuclear infiltrate
		Feature 33: Band-like infiltrate

ب. پیش پردازش

پیش پردازش داده ها گام مهمی در فرآیند داده کاوی می باشد. در داده کاوی روشهای جمع آوری داده معمولا چندان کنترل شده نیستند و مقادیر پرت، تکراری و یا اشتباه می تواند منجر به خروجی نامعتبر شوند. بنابراین انجام مجموعه فعالیت هایی قبل از بهره گیری از داده ها جهت تضمین کیفیت لازم است.[۱۴] اگر اطلاعات دارای افزونگی یا غیر مرتبط در فرآیند استخراج دانش و فرآیند آموزش موجود باشند، این مراحل بسیار مشکل خواهند شد. ممکن است گام پیش پردازش زمانبر باشد و شامل فرآیندهایی از قبیل پاکسازی، نرمالسازی، تبدیل و استخراج ویژگیها باشد. محصول این مرحله مجموعه آموزشی داده خواهد بود.[۱۵] در این پژوهش پیش پردازش داده ها طی دو مرحله انجام شد که نخستین گام حذف داده های

ناقص از مجموعه داده جهت سازگاری بهتر در مراحل بعدی و دومین گام انتخاب ویژگیها براساس استراتژی و الگوریتم Relief-F بود. استراتژی Relief-F نمونه ها را به صورت تصادفی انتخاب می کند و وزن ویژگیها را بر اساس نزدیکترین همسایه محاسبه می نماید. این استراتژی یکی از موفق ترین روشها در انتخاب ویژگیها می باشد.[۱۶] پس از آزمایش مقادیر مختلف ویژگیها و محاسبه خروجی ها، تمام ویژگیها بدون حذف انتخاب شدند و در ورودی الگوریتم اعمال شدند.

ج. الگوریتم کای- نزدیکترین همسایه[۷]

الگوریتم KNN یکی از سهل ترین و مستقیم ترین تکنیک های داده کاوی می باشد. به این روش طبقه بندی مبتنی بر حافظه نیز گفته می شود.[۱۷] جهت کار با ویژگیهای پیوسته و گسسته، اختلاف بین ویژگیها با استفاده از فاصله اقلیدسی محاسبه می شود. برای مثال، اگر نمونه اول عبارت باشد از ($a_1,a_2,a_3,\ldots,a_n$) و نمونه دوم برابر با ($b_1,b_2,b_3,\ldots,b_n$) باشد، فاصله ی بین آنها با استفاده از فرمول ۱ محاسبه می شود:

$$\sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\cdots+(a_n-b_n)^2} \quad (1)$$

مهمترین مشکل در کاربرد فرمول فاصله ی اقلیدسی این است که مقادیر بزرگتر غالبا مقادیر کوچکتر را غرق می کنند. KNN معمولا با داده های پیوسته کار می کند اما برای داده های گسسته نیز کارآمد است. در مقادیر گسسته نمونه های فرضی a_1 و b_1، اگر مقادیر این نمونه ها باهم تفاوت داشته باشند، مقدار اختلاف بین آنها برابر ۱ و در صورت نبود تفاوت معادل صفر در نظر گرفته می شود[۱۰]. الگوریتم به ازای مقادیر مختلف همسایگی تا رسیدن به بالاترین حساسیت و ویژگی اجرا شد که در بخش نتایج به آن اشاره شده است.

د. *K- fold cross validation*

Cross Validation روشی برای تخمین عمومیت خطا بر مبنای نمونه گیری متعدد است و نتیجه چنین تخمینی غالبا برای انتخاب بین مدلهای مختلف از جمله معماری شبکه کاربرد دارد. در K-fold CV داده ها و نمونه ها به k زیر مجموعه تقریبا برابر تقسیم می شوند و سپس شبکه به تعداد k مرتبه مورد آموزش قرار می گیرد و در هر مرتبه یکی از زیر مجموعه ها کنار گذاشته می شوند.[۱۸] برای پژوهش جاری از تقسیم بندی ۱۰ تایی استفاده شد و دقت و حساسیت خروجی ها محاسبه شدند تا در صورت نیاز تغییرات مورد نیاز اعمال گردد. با بررسی بهترین دقت و حساسیت ، ۴۰ در صد از نمونه ها (۳۷ نمونه) جهت نمونه گیری تخصیص داده شدند و مابقی نمونه ها (۳۲۹ نمونه) برای انجام آزمونهای خروجی کنار گذاشته شد. اندازه نسبی مجموعه آموزشی ۸۰ درصد تعیین و نمونه گیری ۱۰ بار تکرار شد.

۴. یافته های پژوهش

جدول شماره ۲ حساسیت، اختصاصی بودن و صحت نتایج الگوریتم کای - نزدیکترین همسایه را در تشخیص بیماری پوستی اریتماتواسکاموز به ازای مقادیر مختلف K نشان می دهد. بالاترین صحت بدست آمده برابربا ۹۸,۰۱ درصد می باشد که به ازای K=5 حاصل شده است. میزان حساسیت و اختصاصی بودن نیز برابر با ۱۰۰ درصد بود.

در مطالعات انجام شده روی این مجموعه داده بالاترین میزان صحت برابر ۰,۹۸۲۳ می باشد که با استفاده از روش ماشین برداری پشتیبان[۸] بدست آمده است.

[۷] K-Nearest Neighbour(KNN)
[۸] Support Vector Machine

[۱۹] برای K=5 تعداد پاسخهای صحیح و اشتباه در جدول ۳ بصورت ماتریس آسیمگی نشان داده شده است.

جدول ۲ – محاسبه کارایی الگوریتم KNN به ازای مقادیر مختلف K

	اختصاصی بودن[۱۱]	حساسیت[۱۰]	صحت[۹]
K=1	1	1	0.9643
K=2	0.9949	0.9882	0.9625
K=3	1	1	0.9750
K=4	1	1	0.9732
K=5	1	1	0.9801
K=6	0.9913	0.9798	0.9714
K=7	1	1	0.9732
K=8	1	1	0.9715
K=9	0.9915	0.9798	0.9768
K=10	1	1	0.9668

جدول ۳ – ماتریس آسیمگی تشخیص ها به ازای K=5

	Psoriasis	Seboreic dermatitis	Lichen Planus	Pityriasis rosea	Chronic dermatitis	Pityriasis Rubra pilaris
Psoriasis	160	0	0	0	0	0
Seboreic dermatitis	0	77	0	2	1	0
Lichen Planus	0	0	100	0	0	0
Pityriasis rosea	0	5	0	55	0	0
Chronic dermatitis	0	2	0	0	68	0
Pityriasis rubra pilaris	0	0	0	0	0	30

این مجموعه داده استاندارد در پژوهش های متعددی در حوزه هوش مصنوعی و یادگیری ماشین مورد استفاده قرار گرفته است که در جدول ۴ روش و برخی الگوریتم های معروف مورد استفاده در یادگیری ماشین و میزان صحت خروجی این روش ها نشان داده شده است.

۵. بحث و نتیجه گیری

در این پژوهش برای داده کاوی روی مجموعه داده استاندارد، رویکردی مبتنی بر الگوریتم کا نزدیکترین همسایه به همراه پیش پردازش بر مبنای اتراتژی Relief-F و نمونه سازی Cross Fold برای تشخیص بیماری اریتماتو اسکاموز پیشنهاد شد. الگوریتم کا نزدیکترین همسایه با پیش پردازش کارآمد و انتخاب همسایگی صحیح عملکرد مناسبی در طبقه بندی داده های پزشکی دارد. میزان دقت و حساسیت ۱۰۰ درصدی این رویکرد و همچنین صحت فراتر از ۹۸ درصد، این رویکرد را در کنار روشهای نوین هوش مصنوعی به عنوان الگوریتمی کم هزینه و پر کاربرد قرار داده

[9] Accuracy
[10] Sensitivity
[11] Specificity

است. انتخاب ابزار مناسب جهت پیاده سازی روشهای هوش مصنوعی به خصوص یادگیری ماشین منجر به سهولت در طراحی و عملیاتی نمودن سیستم های خبره و تصمیم یار در حیطه سلامت خواهد شد.

جدول ۴. میزان صحت روش مورد استفاده ما و سایر روشهای انجام شده طبقه بندی در مطالعات

عنوان پژوهش	روش و الگوریتم	میزان صحت طبقه بندی
Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems[19]	ANFIS	**0.9550**
Similarity classifier with generalized mean applied to medical data[1]	Fuzzy SBC	**0.9702**
An ensemble of classifiers for the diagnosis of erythemato-squamous diseases[3]	LSVM	**0.9722**
	RS	**0.9722**
	B2-15	**0.9830**
The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-NN based weighted pre-processing methods to diagnosis of erythemato-squamous diseases[20]	Fuzzy weighted	**0.8818**
Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases[21]	IFSFS و SVM	**0.9861**
روش مورد مطالعه ما	KNN with preprocessing	**0.9801**

REFERENCES

1 Luukka P, Leppälampi T. Similarity classifier with generalized mean applied to medical data. Computers in Biology and Medicine, 2006; 36(9): 1026-1040.

2 Güvenir HA, Emeksiz N. An expert system for the differential diagnosis of erythemato-squamous diseases. Expert Systems with Applications, 2000; 18(1): 43-49.

3 Nanni L. An ensemble of classifiers for the diagnosis of erythemato-squamous diseases. Neurocomputing, 2006; 69(7–9): 842-845.

4 Giveki D. Detection of erythemato-squamous diseases using AR-CatfishBPSO-KSVM. Signal & Image Processing, 2012; 2 (4): 57-72.

5 Han J. Data Mining: Concepts and Techniques. 2005; Morgan Kaufmann Publishers Inc.

6 Lee IN, Liao SC, Embrechts M. Data mining techniques applied to medical information. Med Inform Internet Med, 2000; 25(2):81-102.

7 Obenshain, MK. Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology, 2004; 25(8): 690-695.

8 Liao SC, Lee IN. Appropriate medical data categorization for data mining classification techniques. Informatics for Health and Social Care, 2002; 27(1): 59-67.

9 Liao SC, Lee IN. Appropriate medical data categorization for data mining classification techniques. Med Inform Internet Med, 2002; 27(1): 59-67.

10 Shouman M, Turner T, Stocker AR. Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients. International Journal of Information and Education Technology, 2012; 2(3).

11 Athitsos V. Boosting nearest neighbor classifiers for multiclass recognition, in Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on, 2005: IEEE: 45.

12 Moreno-Seco F, Micó L, Oncina J. A modification of the LAESA algorithm for approximated k-NN classification. Pattern Recognition Letters, 2003; 24(1–3): 47-53.

13 Frank A, Asuncion A. UCI Machine Learning Repository, University of California, 2010.

14 Pyle D. Data preparation for data mining. 1999: Morgan Kaufmann Publishers Inc.

15 Kotsiantis SB, Kanellopoulos D, Pintelas PE. Data Preprocessing for Supervised Leaning. International Journal of Computer Science, 2006; 2(1).

16 Liu H, Motoda H. Computational Methods of Feature Selection. 2007: Taylor & Francis.

17 Alpaydin E. Voting over Multiple Condensed Nearest Neighbors. Artif. Intell. Rev., 1997; 11(1-5): 115-132.

18 Sarle W. What are cross-validation and bootstrapping?, 2012; Available from: http://www.faqs.org.

19 Übeylı ED, Güler I. Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems. Computers in Biology and Medicine, 2005; 35(5): 421-433.

20 Polat K, Gunes S. The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-NN based weighted pre-processing methods to diagnosis of erythemato-squamous diseases. Digit. Signal Process., 2006; 16(6): 922-930.

21 Xie J, Wang C. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. Expert Syst. Appl., 2011; 38(5): 5809-5815.