1ST National Conference on
Applications of Intelligent Systems(Soft Computing)
in Science and Technology
Islamic Azad University,Quchan Branch
4 -5 ,March,2013

اولین همایش ملی
کاربرد سیستم های هوشمند(محاسبات نرم) در علوم و صنایع
دانشگاه آزاد اسلامی واحد قوچان

یک رویکرد جدید برای تشخیص کم کاری تیروئید با استفاده از تکنیک های داده کاوی

احسان یوسف زاده[۱]، ابوالفضل کاظمی [۲]،پرهام عظیمی[۳]

۱ -دانشجوی کارشناسی ارشد، دانشکده مهندسی صنایع و مکانیک،دانشگاه آزاداسلامی،واحدقزوین،قزوین،ایران

۲ -استادیار، دانشکده مهندسی صنایع و مکانیک،دانشگاه آزاداسلامی،واحدقزوین،قزوین،ایران

۳ - استادیار،دانشکده مهندسی صنایع و مکانیک،دانشگاه آزاداسلامی،واحدقزوین،قزوین،ایران

چکیده

موضوع:انتخاب ویژگی یکی از مهمترین گام های داده کاوی می باشد، چرا که به طور مستقیم بر نتایج بدست آمده اثر می گذارد.انتخاب ویژگی ابعاد مساله را کاهش ومیزان دقت مدل را افزایش می دهد.همچنین الگوهای بدست آمده قابل فهم تر می شوند.تکنیک های زیادی برای انتخاب ویژگی وجود دارد. وزن دهی به ویژگی ها یکی از تکنیک های انتخاب ویژگی می باشد.

هدف:هدف تحقیق انتخاب ویژگی های مهم تیروئید وتحلیل این داده ها با بالاترین دقت ممکن می باشد.

روش تحقیق:در این تحقیق یک رویکرد جدید برای تحلیل داده های تیروئید ارائه گردید.انتخاب ویژگی به وسیله روش وزن دهی مهمترین آنالیز جزئی(PCA[4])انجام گرفت.تحلیل تیروئید براساس ویژگی های انتخاب شده توسط روش نزدیکترین همسایگی انجام گرفت.

نتیجه:دقت رویکرد انجام گرفته ۱۰۰% می باشد.براساس بهترین دانش ما این رویکرد در تحلیل داده های تیروئید هنوز به کار نرفته است.

کلمات کلیدی:تیروئید،انتخاب ویژگی،روش مهم ترین آنالیز جزئی،نزدیکترین همسایگی

۱ -معرفی

در دنیای واقعی، داده ها معمولادارای مقادیر مفقوده،نویز دار وناسازگار می باشند.همچنین داده ها دارای ویژگی های اضافی هستند که بهبودی در نتایج حاصل نمی کنند ودر اکثر مواقع کیفیت نتایج را خراب می کنند. بنابراین با حذف

[۱] دانشجوی کارشناسی ارشد،e.yousefzadeh@qiau.ac.ir، ۰۹۱۹۶۱۶۶۲۳۰

[۲] استادیاردانشگاه،abkaazemi@qiau.ac.ir

[4]Analysis component principle

ویژگی های اضافی دقت روشهای دسته بندی کننده بیشتر می شود(۱).در داده های پزشکی ویژگی های اضافی نیز باعث سخت تر شدن مساله و بالارفتن ابعاد مساله می شوند، که این ویژگی ها باید حذف شوند(۲).

اکثر تکنیک ها و مدلهای داده کاوی که برای پیش بینی و دسته بندی به کار می روند، دارای خطا می باشند که ناشی از وجود ویژگی های اضافه در مساله می باشد.مادر این تحقیق سعی کردیم رویکردی از انتخاب ویژگی و روشی از دسته بندی ارائه کنیم که دارای خطای صفر باشد.

۲ -روش

۱.۲ -داده

داده های تیروئید از سایت معتبر Uci Repository گرفته شده است.تعداد داده های این بیماری ۲۱۵ می باشد.این داده دارای پنج ویژگی و یک ویژگی کلاس با سه حالت می باشد(جدول۱).هدف این داده تعیین وضعیت بیماران تیروئیدی می باشد.

نام ویژگی
Class (1 = normal, 2 = hyper, 3 = hypo)
T3-resin uptake test. (A percentage)
Total Serum thyroxin as measured by the isotopic displacement method.
Total serum triiodothyronine as measured by radioimmunoassay.
Basal thyroid-stimulating hormone (TSH) as measured by radioimmuno assay.
Maximal absolute difference of TSH value after injection of 200 micro grams thyrotrophic-releasing hormone as compared to the basal value.

۲.۲ -روشهای داده کاوی

PCA- ۱.۲.۲

PCAیکی از الگوریتم های هدایت نشده می باشد که برای کاهش ویژگی ها به کار می رود.وزن دهی به ویژگی ها یکی از مکانیزم های PCA می باشد که در نهایت مهمترین ویژگی ها را بدست می آورد.اما مهمترین مکانیزم PCAترکیب ویژگی ها به صورت مناسب باهم می باشد،به گونه ای که حداکثر خواص ویژگی های اولیه حفظ گردد (۸).

۲.۲.۲-نزدیکترین همسایگی

این روش با استفاده از نمونه های قبلی که در اختیار دارد،یک نمونه جدید را به نزدیکترین نمونه از لحاظ شباهت اختصاص می دهد.به عبارت دیگر کلاس نمونه جدید همان کلاس نزدیکترین نمونه از لحاظ شباهت به نمونه فوق می باشد(۹).

۳.۲.۲-انتخاب ویژگی

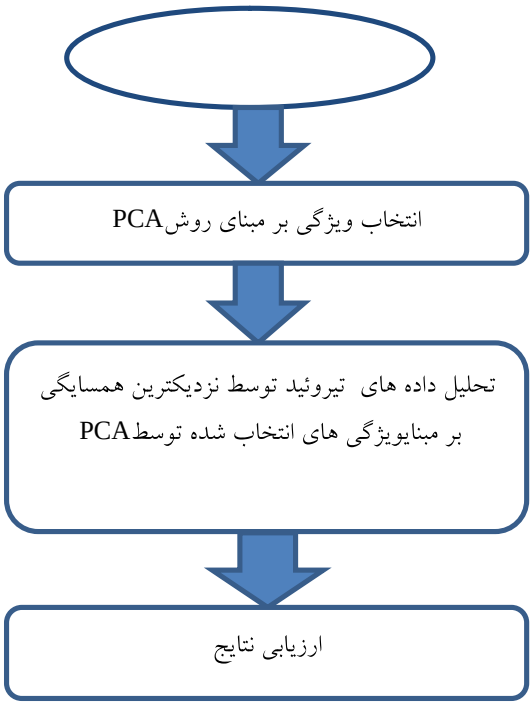
انتخاب ویژگی مکانیزمی است که یک سری ویژگی از کل تعداد ویژگی ها را براساس یک سری معیار انتخاب می کند به گونه ای که کیفیت نتایج بدست آمده نسبت به حالت بدون حذف ویژگی مساوی یا بهتر شود(۳و٤و٥).هدف انتخاب ویژگی کاهش هزینه،کاهش ابعاد و پیچیدگی مساله،بالا بردن دقت مدل و قابل فهم کردن نتایج می باشد.(٦و٧).

۳.۲-توصیف روش

تحلیل داده های تیروئید بر مبنای PCAونزدیکترن همسایگی می باشد(شکل۱).در ابتدا، برای تعیین ویژگی های مهم وتاثیر گذار برای تیروئید وهمچنین بالا بردن دقت روش داده ها به روش PCAداده شده است.این روش با مکانیزمی که اعمال می کند، به هر ویژگی یک وزن بین عدد صفر و یک داده است(جدول۳).

مطابق با خروجی PCA، ویژگیSerum triiodothyronine وزنی نگرفته است(ویژگی بی اهمیت).بنابراین این ویژگی از مجموعه ویژگیها حذف و سایر ویژگی ها به روش نزدیکترین همسایگی داده شده است.

روش نزدیکترین همسایگی پیش بینی بیماری تیروئید بر اساس ویژگی های گرفته شده را انجام داده است.



شکل۱-فلوچارت کلی مدل

Serum triiodothyronine	۰
TSH	۰.۱٦۰۹
Serum thyroxin	۰.۱٦۸
Maximal absolute difference of TSH	۰.۲٦۳۳
T3-resin	۱

۳-نتایج

مدل توسط نرم افزار Rapid Miner5 و بر روی یک لب تاپ با مشخصات COREi5 و 4G RAM اجرا شده و نتایج آن به صورت زیر می باشد.

۱-مطابق با آنچه گفته شد،براساس رویکرد ارائه شده، صحت پیش بینی بیماری تیروئید ۰۰% می باشد.این نتیجه براساس جدول ۳و٤ قابل مشاهده است.

جد

دقت	۰۰%
حساسیت	۰۰%
مشخصه	۰۰%

۰	۰	۱۵۰
۰	۳۵	۰
۳۰	۰	۰

۲-دقت مدل بر اساس ٤ ویژگی نهایی تیروئید بدست آمده است(جدول ۵).

نام ویژگی ها
TSH
Serum thyroxin
Maximal absolute difference of TSH
T3-resin

۳-نتیجه رویکرد ارائه شده با تحقیق(۱۰) مقایسه گردیده است(جدول٦).

مرجع	نوع روش	دقت
این تحقیق	رویکرد ارائه شده	%۱۰۰
Wei-Wen	سیستمهای ایمنی مصنوعی	%۹۸.۳۹

۱.۳ -بحث وکار آینده

در این تحقیق یک رویکرد جدید برای تحلیل تیروئید ارائه گردید.انتخاب ویژگی توسط روش PCA انجام و داده های تیروئید توسط روش نزدیکترین همسایگی و بر اساس ویژگی های انتخاب شده انجام گرفت.دقت این رویکرد ۱۰۰%(خطای صفر)می باشد.با توجه به نتیجه بدست آمده و مقایسه خروجی با مقاله انجام شده،کارایی رویکرد ارائه شده، قابل توجه می باشد.

از این رویکرد برای سایر موضوعات پزشکی و غیر پزشکی می توان استفاده نمود.همچنین به جای PCA برای وزن دهی ویژگیها، می توان از سایر روشها مانند ماشین بردار پشتیبان برای وزن دهی ویژگی ها استفاده نمود.

مراجع

1-K.C. Tan a, E.J. Teoh a, Q. Yua, K.C. Goh a.2009. A hybrid evolutionary algorithm for attribute selection in data mining, Expert Systems with Applications, Expert Systems with Applications 36: 8616:8630.
2-Sean N. Ghazavi, Thunshun W. Liao.2008. Medical data mining by fuzzy modeling with selected features, Artificial Intelligence in Medicine 43:195:206.
3-Liu, H., &Motoda, H. 1998. Feature selection for knowledge discovery and data mining. Kluwer Academic Publishers
4-Pena, J. M., Lozano, J. A., Larranaga, P., &Inza, I. 2001.Dimensionality reduction in unsupervised learning of conditional gaussian networks. IEEE Transactions onPattern Analysis and Machine Intelligence, 23(6): 590:603
5-Yu, L., & Liu, H. 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th internationalconference on machine learning (pp. 856:863).
6-Witten, I. H., & Frank, E. 2000. Data mining: practical machine learning tools andtechniques with Java implementations. San Francisco: Morgan Kaufman.

7-Hall, M. A. 1999. Correlation-based subset feature selection for machine learning.Waikato: University of Waikato.
8-X. Xu, X.N. Wang. 2005. An adaptive network intrusion detection method based on PCA and support vector machines, Lecture Notes in Artificial Intelligence. 3584: 696:703.
9-Jeng, B. C., & Liang, T. P. 1995.Fuzzy indexing and retrieval in case-based system.Expert Systems with Applications, 8(1): 135:142.
10-Wei-Wen. Chang a, Wei-Chang Yeh a, Pei-Chiao Huang b,.2010.A hybrid immune-estimation distribution of algorithm for mining thyroid gland data, Expert Systems with Applications .37:2066:2071.

A novel approach for low working thyroid by using data mining techniques.

Abstract

Subject: Feature selection is one of the most steps in data mining. Because it affects on obtained results directly. Feature selection reduces problem dimensions and increases model accuracy rate. Also, attained patterns will be more understandable. There are many techniques for feature selection. Giving weight to features is one of technique for feature selection.
Goal: The goal of this paper is, selecting the most important thyroid features and analyzing these data with the highest accuracy as possible as.
Research method: In this paper, a novel approach was offered for analyzing gland thyroid data. Feature selection was executed by analysis component principle (PCA) method.Analysis gland thyroid data was executed on based selected features with K-nearest neighborhood method.
Results: Executed approach accuracy is100% .On bests our knowledge; this approach hasn't been done on gland thyroid data yet.

Key words: Thyroid, Feature selection, Analysis component principle, K-nearest neighborhood.