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

بهینه‌سازی یک سیستم خبره توسط الگوریتم ژنتیک به منظور پیش‌بینی هپاتیت

نوید عربی۱، نیلوفر راستین۲، دکتر شهرام جعفری۳

۱ - دانشجو کارشناسی ارشد دانشگاه شیراز ، دانشکده مهندسی برق و کامپیوتر

۲ - دانشجو کارشناسی ارشد دانشگاه شیراز ، دانشکده مهندسی برق و کامپیوتر

۳- استادیار دانشگاه شیراز ، دانشکده مهندسی برق و کامپیوتر

چکیده

تشخیص بیماری هپاتیت با توجه به علائم بالینی اندکی که در مراحل اولیه دارد، بسیار دشوار است. روش‌هایی که بتوانند بر اساس استخراج دانش از داده های پزشکی به تشخیص زود هنگام این بیماری کمک کنند، بسیار حائز اهمیت هستند. یکی از روش‌های مطرح در استخراج دانش، استفاده از الگوریتم های داده کاوی[4] است.که این روش ها در مقایسه با روش‌های یادگیری ماشین، دقت کمتری در پیش بینی دارند. این روش‌ها مجموعه قوانینی از داده ها را استخراج می‌کنند که قابل تجزیه و تحلیل می‌باشد. در نتیجه در مواردی که نیاز به استنتاج داریم، از جمله موارد پزشکی نقش بسیار مهمی ایفا می‌کنند. در این مقاله ابتدا با استفاده از الگوریتم درخت تصمیم[5] و روش یادگیری قانون وابستگی[6] که دو روش رایج و قدرتمند داده کاوی هستند قوانین متعددی از مجموعه داده‌های مربوط هپاتیت استخراج شده و سپس یک زیر مجموعه از موثرترین این قوانین توسط الگوریتم ژنتیک[7] انتخاب گردیده است. دقت نتایج حاصله با روش کلاسه بندی نایوبیز[8] مورد مقایسه قرار گرفته ودقت ۷۷%را نشان داده است. نتیجه نمایانگر ٦% افزایش دقت نسبت به روش کلاسه بندی نایوبیز می‌باشد. بر اساس مطالب گفته شده، یک سیستم خبره پیش بینی

۱ دانشجو، ۰۹۱۳۱۹۹۵۳۹۷ – navidved@gmail.com

۲ دانشجو ، 09364517379 – niloofar.rastin@gmail.com

3 استادیار، jafaris@shirazu.ac.ir

[4] Data Extraction
[5] Decision Tree
[6] Association Rule Mining
[7] Genetic Algorithm
[8] Naive Bayes Classifier
[9] Clips Language Programming

کننده توسط نرم افزار کلیپس[9] طراحی شده است، که می‌تواند در امر پیش بینی بیماری هپاتیت به افراد متخصص این حوزه کمک بالقوه‌ای کند.

کلید واژه : پیش‌بینی هپاتیت، الگوریتم داده کاوی، الگوریتم درخت تصمیم، روش یادگیری قانون وابستگی، االگوریتم ژنتیک، روش کلاسه بندی نایو بیز ، سیستم خبره.

Genetically Optimized Rule based Expert System for Hepatitis Prediction

N. Arabi ; N. Rastin ; S. Jafari
Electrical and Computer Engineering
Shiraz University, Shiraz, Iran

Abstract

Diagnosis of Hepatitis is extremely difficult, due to few clinical symptoms in its early stages. Methods enabling the extraction of knowledge from medical data to help early diagnosis of this disease are of critical importance. One of the methods for extracting knowledge is using data extraction algorithms. In general, these methods are less accurate in their prediction comparing to those of machine learning techniques. Data extraction methods extract sets of rules which can be analyzed by human being. Therefore, in cases which are in need of inference such as medical cases it plays an important role. In this dissertation two popular and powerful data extraction methods are used which are Design Tree Algorithm and Associative rule mining method. many rules have been extracted from hepatitis related set and then, a subset of the most effective rules will be selected by means of genetic algorithm. Accuracy of the results were compared according to Naïve bayes classification method showing 77% accuracy which representing 6% increase of accuracy comparing to the Naïve bayes classification method. Based on the knowledge obtained from the previous steps a predictive expert system designed by clips could potentially help to predict the hepatitis disease by related experts in this area.

Keywords

Hepatitis Prediction, Data Extraction , Decision Tree, Associative Rule Mining, Genetic Algorithm, Naive Bayes Classifier, Expert System.

۱ - مقدمه

بیماری هپاتیت یک بیماری کبدی است که در مراحل اولیه علائم بالینی اندکی دارد و بدین دلیل تشخیص زود-هنگام آن بسیار دشوار است. روشی که بتواند این بیماری را در مراحل اولیه آن پیش بینی کند می تواند بسیار ارزشمند باشد و باعث کاهش عوارض احتمالی بعدی آن شود. در حال حاضر دقیق‌ترین روش تشخیص این بیماری آزمایش خون می‌باشد که با توجه به بالینی نبودن علائم این بیماری، افراد کمتر به انجام این آزمایش مبادرت می‌کنند. بنابراین طراحی و ساخت یک سیستم خبره با هدف پیش بینی بیماری هپاتیت، که بر پایه برخی خصوصیات پزشکی عمل می کند، بسیار کارآمد خواهد بود. با توجه به این واقعیت که قواعد صریحی برای طراحی این سیستم خبره در

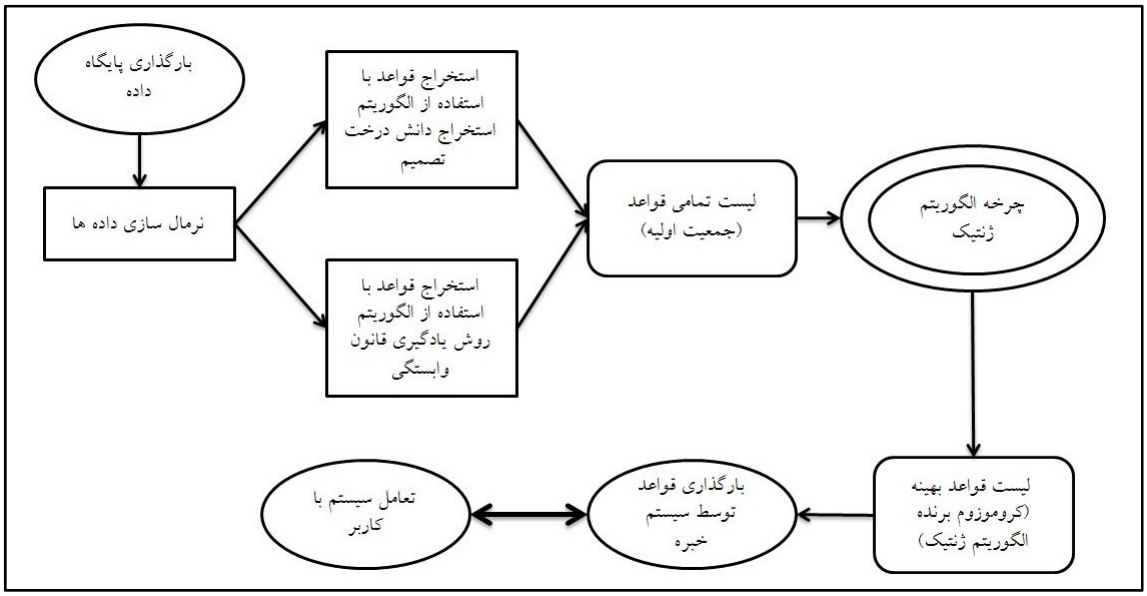
دسترس نیست، بهترین راهکار استخراج قواعد از داده‌های پزشکی مرتبط با ایـن بیمـاری مـی‌باشـد. بـدین منظـور از روش‌های استخراج دانش از داده، برای به‌دست آوردن این گونه قواعد استفاده مـی‌کنـیم. تحقیقـات مشـابهی در ایـن رابطه انجام شده است. در [1] برای پیش بینی هپاتیت از یک سیستم خبره بر پایه حقایق فازی استفاده کرده که توسط یک شبکه‌ی عصبی بهینه شده است. در [2] از روش انتخاب خصوصیات برای استخراج ویژگی های مهـم داده هـا و روش یادگیری ماشین بردار پشتیبانی برای کلاسه بندی استفاده شده است.

در این مقاله از درخت تصمیم و یادگیری قانون وابستگی که روش‌های قدرتمند و رایج استخراج دانـش هسـتند، برای استخراج قواعد از داده‌های هپاتیت [3] استفاده شده و دقت هر الگوریتم به صورت جداگانه توسط روش کلاسه بندی نایو بیز مورد سنجش قرار گرفته است. برای بهتر شدن نتایج، قواعد استخراج شده به صورت کروموزوم ژنتیکی تعریف شده، سپس با استفاده از قدرت بهینه سازی الگوریتم ژنتیک ، بهترین قواعد، یعنی آن‌هایی که بیشترین دقت را دارند استخراج گردیده‌اند. نهایتا قواعد بهینه شده توسط زبان برنامه‌نویسی کلیپس به یک سیستم خبـره قابـل اسـتفاده توسط افراد متخصص این حوزه تبدیل شده است.

.

۲- روش بررسی

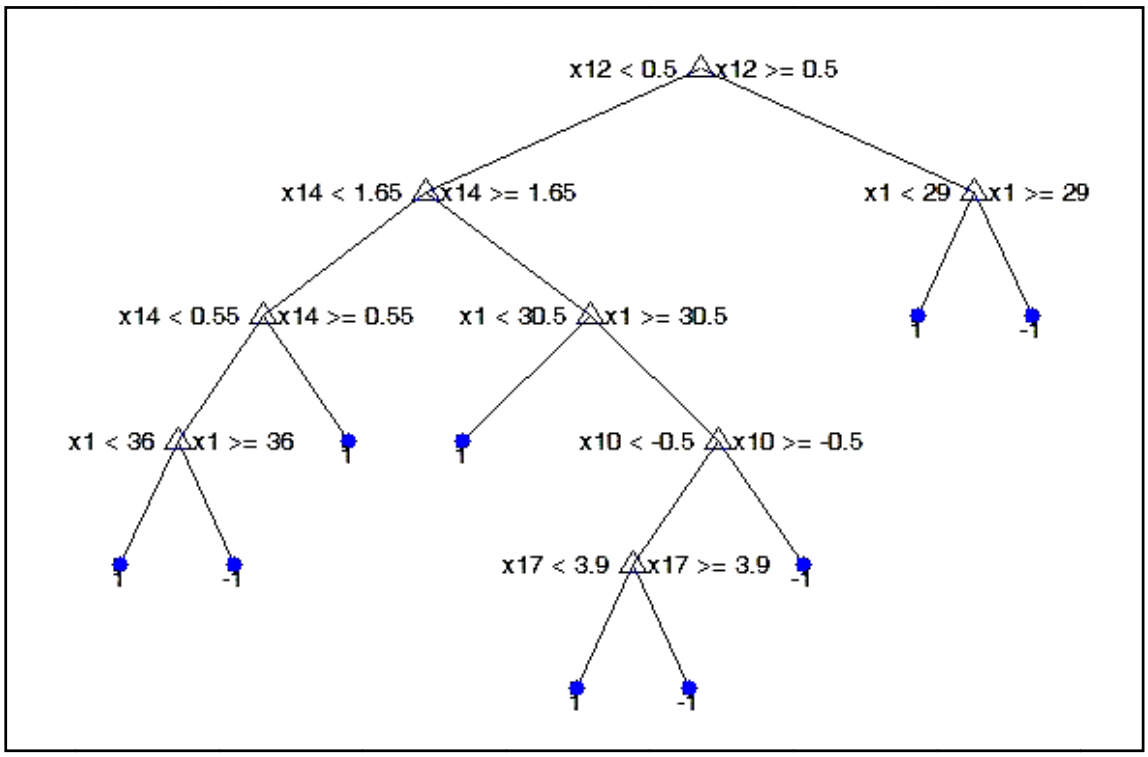
ابتدا برای حذف اثری که واحدهای اندازه گیری متفاوت در نتایج می‌گذارند و نیز متناسب کردن مقیاس داده ها ، داده های اصلی نرمال سازی شده و از آن پس داده‌های نرمال شده در تمامی مراحل فرآیند استفاده شده است. در ادامه مراحل فرآیند (شکل شماره ۱) به تفصیل آورده شده.



شکل شماره ۱ - فرآیند استخراج دانش

۲. ۱- الگوریتم استخراج دانش درخت تصمیم

الگوریتم درخت تصمیم روشی برای تقریب توابع هدف با مقادیر گسسته است. این روش نسبت به نـویز داده‌هـا مقاوم بوده و قادر است ترکیب فصلی گزاره های عطفی را یاد بگیرد. این روش یکـی از مشـهورترین الگـوریتم‌هـای یادگیری استقرائی است که در کاربردهای مختلف به‌طور موفقیت آمیز به کار گرفته شده است. درخت تصمیم درختی است که در آن نمونه‌ها به نحوی دسته‌بندی می‌شوند که از ریشه به سمت پائین رشد کنند و در نهایـت بـه گـره‌هـای برگ برسند. هر گره داخلی یا غیر برگ با یک ویژگی مشخص می‌شود. ما بـا اسـتفاده از داده هـای نرمـال شـده و بـا استفاده از الگوریتم درخت تصمیم یک مجموعه قوانین بدست آوردیم (شکل شماره ۲) . هر مسـیر ریشـه تـا بـرگ های این درخت نماینده یک قاعده استخراج شده می‌باشد. در این درخت تصمیم ۱+ و ۱- واقع در هر برگ درخـت نماینده دو کلاس"بیمار بودن" و "بیمار نبودن" است. قوانین استخراج شده به صورت یک سیستم خبره پیاده سـازی شد با استفاده از داده های تست ، دقت ٤٧٪ حاصل بدست آمد.

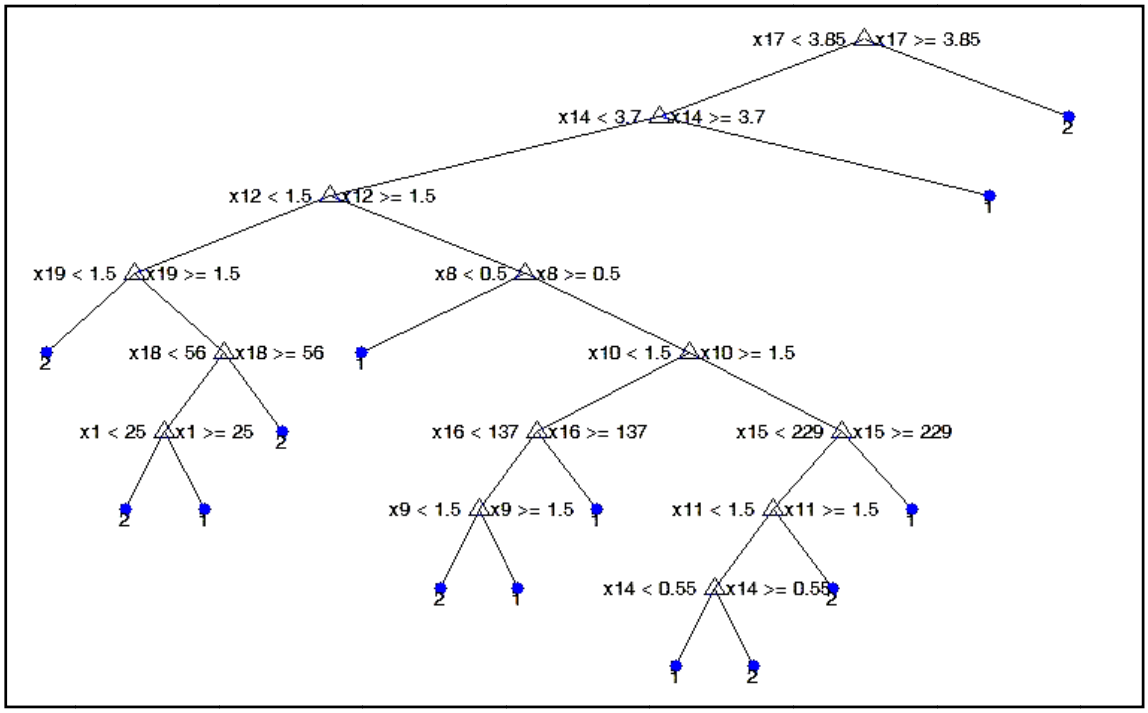


شکل شماره 2 - درخت تصمیم بدست آمده از الگوریتم درخت تصمیم

۲. ۲- روش یادگیری قانون وابستگی

در علم داده‌کاوی، روش یادگیری قانون وابستگی یک روش بهینه برای یافتن قواعد صریح بین متغیرهای موجـود در پایگاه داده ها است. در [4] چگونگی تحلیل و ارائه این قوانین و روابط یافته شده در پایگـاه‌هـای داده بـا اسـتفاده از معیارهای متفاوت سنجش درستی توضیح داده شده است. در [5] بر مبنای مفهوم قوانین صریح، قـوانین وابسـتگی را برای کشف قاعده های موجود در داده‌های تراکنشی با مقیاس بالا معرفی گردیده است.

در این مقاله دو روش بالا مورد استفاده قرار گرفته و با استفاده از داده های نرمال شده و الگوریتم یادگیری قانون وابستگی یک درخت تصمیم به‌دست آمده است (شکل شماره 3) . با استخراج و پیاده سازی توسط زبان برنامه-نویسی کلیپس، سیستم خبره ای بر پایه قواعد به‌دست آمده پیاده‌سازی شده و با استفاده از داده‌های تست ، دقت ٦٣٪ حاصل گردیده است. در این درخت تصمیم اعداد ۱ و ۲ نماینده‌ی کلاس‌های "بیمار بودن" و "بیمار نبودن" می-باشند.



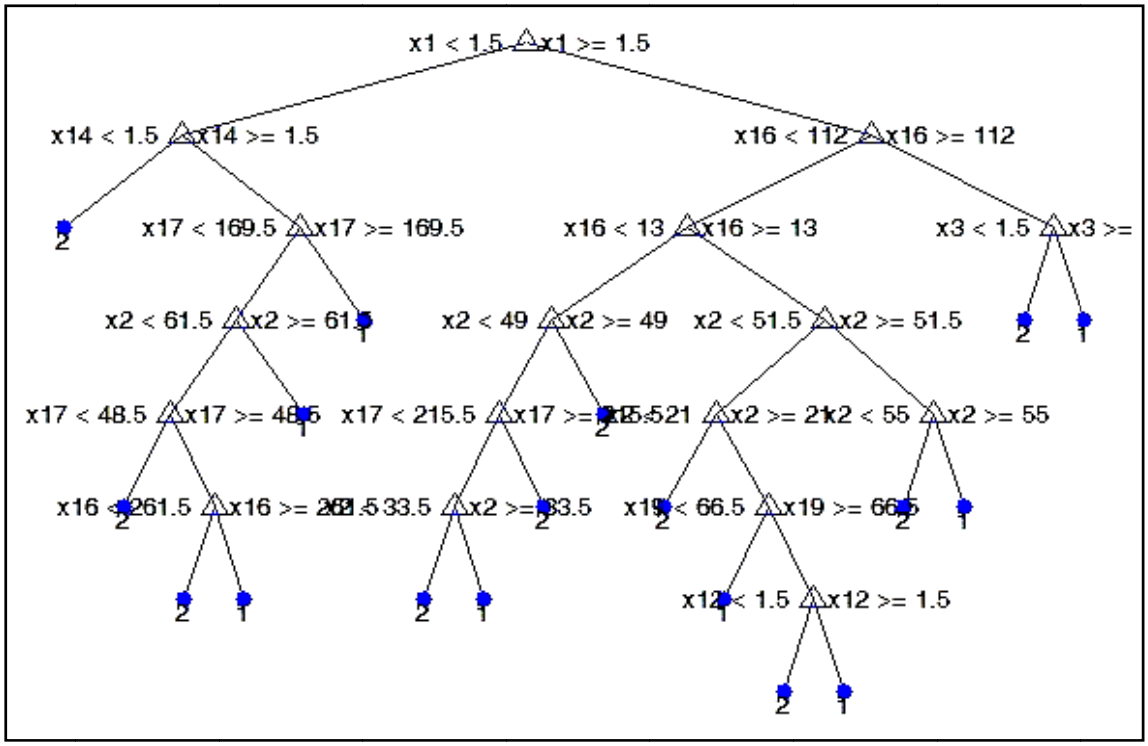
شکل شماره 3 - درخت تصمیم بدست آمده از الگوریتم یادگیری قانون وابستگی

۲. ۳- بهینه سازی قواعد استخراج شده در دو قسمت قبل توسط الگوریتم ژنتیک

در این مرحله از یک الگوریتم ژنتیک به عنوان یک بهینه‌ساز قوی برای انتخاب بهترین قواعد از مجموعه قواعد به-دست آمده مراحل قبل مورد استفاده قرار گرفته است. با توجه به ساختار الگوریتم ژنتیک ابتدا ژنوتیپ به صورت یک آرایه ۱ در n تعریف شده است. n تعداد کل قواعد مستقل حاصل شده از دو روش داده‌کاوی مراحل قبل و شماره هر خانه نماینده شناسه یک قانون به‌دست آمده می‌باشد . مقدار هر خانه این آرایه می‌تواند شامل یکی از دو عدد ۰ یا ۱ باشد که نشان دهنده در نظر گرفتن و یا در نظر نگرفتن قاعده اشاره شده توسط این خانه ، در جواب نهایی کروموزوم است. جمعیت اولیه به صورت تصادفی تولید شده و برای همگذری از همگذری یک نقطه‌ای[10] و برای جهش[11] از جهش درجی[12] استفاده شده است. همچنین برای توقف الگوریتم تعداد تکرارها مورد محاسبه قرار گرفته است. تابع

[10] One point CrossOver
[11] Mutation
[12] Insertion mutation

میزان تناسب[13] دقت مجموع قوانین هر کروموزوم می‌باشد که با اعمال داده‌های تست محاسبه گردیده است. تابع انتخاب والدین[14] به این صورت محاسبه شده که ابتدا ۱۰٪ از کل جمعیت را انتخاب و سپس محاسبه میزان تناسب ۲ تا از بهترین را برای اعمال همگذری و جهش گزینش می‌کند. همگذری و جهش هر کدام به ترتیب به احتمال‌های ۱۰۰٪ و ۳۰٪ اتفاق می‌افتد و به صورت پیش فرض برای مشخص کردن حداکثر تعداد دفعات تکرار عدد ۱۰۰۰۰ در نظر گرفته شده است. جواب نهایی کروموزومی است که بیشترین میزان تناسب را دارد. یا به عبارت دیگر جواب نهایی قیودی هستند که بشترین دقت را از داده‌های تست حاصل کرده‌اند . در این مرحله با استفاده از کروموزوم نهایی، یک درخت تصمیم بدست آمده است (شکل شماره 4) . در این درخت تصمیم دو کلاس ۱ و ۲ به نمایندگی "بیمار بودن" و "بیمار نبودن" معرفی شده اند. با استفاده از داده‌های تست دقت مجموعه قوانین آزموده و دقت ۷۷٪ حاصل شده است.



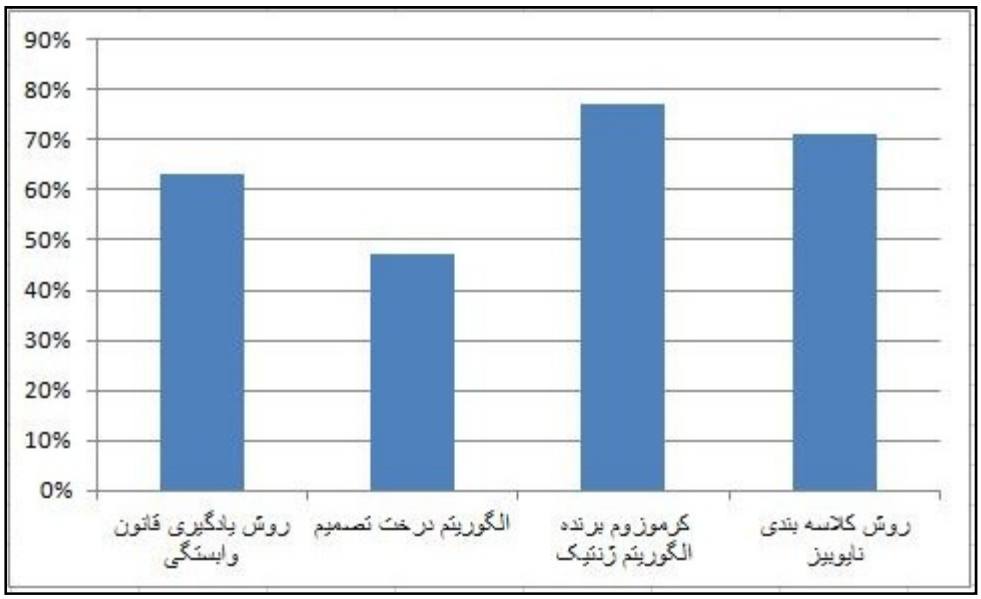
شکل شماره 4 - درخت تصمیم بدست آمده از کروموزوم برنده الگوریتم ژنتیک

۳- نتیجه گیری

جواب نهایی کروموزومی می‌باشد که بیشترین دقت را از داده‌های تست به‌دست آورده است . به عبارت دیگر کروموزوم برنده که پس از اتمام کار الگوریتم از بیشترین میزان تناسب برخوردار است، شماره قوانینی را نشان می‌دهد که در کنار هم بیشترین دقت را به‌دست آورده‌اند. در نتیجه‌ای که به دست آمده این دقت برای داده های تست ۷۷٪ و یا به عبارتی ٦٪ بالاتر از روش کلاسه‌بندی نایو بیز است که نتیجه قابل قبولی می‌باشد (شکل شماره

[13] Fitness Function
[14] Parent Selection Function

5). قواعد حاصل شده توسط زبان برنامه‌نویسی کلیپس به صورت یک سیستم خبره پزشکی آماده سازی شده که قابل استفاده توسط افراد متخصص این حوزه می‌باشد.



شکل شماره ۵ - مقایسه نتایج روش های مختلف استخراج دانش

۵ - مراجع

1- CruzKemal Polat , Salih Gunes. 2007. Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection. Sciencedirect Expert Systems with Applications, Page 484–490.

2- M.Neshat, M.Yaghobi. 2009. Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparing it with Adaptive Neural Network Fuzzy System, WCECS ,Page 133-140.

3- Hepatitis dataset, UCI Machine Learning Repository http://archive.ics.uci.edu/ml/datasets/ Irvine. CA University of California, School of information technology and computer science.

4- Piatetsky Shapiro. 1991. Discovery analysis and presentation of strong rules, AAAI/MIT Press, Cambridge, MA.

5- Rakesh Agrawal. 1993. Mining association rules between sets of items in large databases. ACM SIGMOD international conference on Management of data. Pages 207-216.