Discovering Hidden Cluster Structures in Patients with Cirrhosis Based on Laboratory Data

Mina Pazouki[1], Mohammad Mehdi Sepehri[2], Mehdi Saberifiroozi[3]

[1] MSc, Department of Industrial Engineering, South branch Islamic Azad University, Tehran, Iran.
[2] Associate Professor, Department of Industrial Engineering, Faculty of Engineering, Tarbiat Modares University, Tehran, Iran
[3] Professor, Digestive Disease Research Center, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran.

ABSTRACT

Background:
Liver cirrhosis was one of the most important liver diseases. Other chronic liver diseases could be lead to liver cirrhosis. Liver cirrhosis could be lead one kind of liver cancers named hepatocellular carcinoma. Cirrhosis in the early stages just by laboratory and imaging testes could be diagnosed. In this study cirrhotic patients were classified based on laboratory symptoms. For this purpose data mining approach has been used in this research. Data mining was an interdisciplinary science that discovers the hidden knowledge in the data.

Materials and Methods:
We used K-Means algorithm to categorize the statues of cirrhotic patients. In order to determine the quality of clustering results and to find the best number of clusters, we have used silhouette indices.

Results:
Our data consists of 410 records which have been collected from Dr. Shariati hospital The number of features in this study are .1 items and sampling were divided into two main groups.

Conclusion:
At first, we have done clustering based on 21 attributes and the average silhouette was 41 percent. We improved the model, in order to reach a reasonable structure. Finally, based on 7 attributes, a reasonable clustering model was derived. The new model provides 64 percent average silhouette, and based on patients' status, patients are divided into 2 main categories. The risk of HCC in the first cluster is 23 percent and in the second cluster is 14 percent.

Keywords: Liver cirrhosis; Data mining; clustering; K-means; Silhouette

please cite this paper as:
Pazoki M, Sepehri MM, Saberifiroozi M. Discovering hidden cluster structures in patients with cirrhosis based. *Govaresh* 2014;18:191-7.

Corresponding author:
Mohammad Mehdi Sepehri, MD
Department of Industrial Engineering, Tarbiat Modares University (TMU), Jalal-e Al-e Ahmad Highway, Tehran 1411713114, Iran
Telefax: + 98 21 82883379
E-mail: mehdi.sepehri@gmail.com
Received: 05 May 2014
Edited: 11 Aug. 2014
Accepted: 12 Aug. 2014

مقاله پژوهشی

کشف ساختارهای خوشه‌ای پنهان در بیماران مبتلا به سیروز کبدی بر پایه نشانه‌های آزمایشگاهی

مینا پازوکی[۱]، محمد مهدی سپهری[۲]، مهدی صابری فیروزی[۳]

[۱] دانشجوی کارشناسی ارشد ، دانشکده مهندسی صنایع، دانشگاه آزاد اسلامی واحد تهران جنوب، تهران، ایران
[۲] دانشیار، بخش مهندسی صنایع، دانشکده فنی و مهندسی ، دانشگاه تربیت مدرس، تهران ، ایران
[۳] استاد، مرکز تحقیقات گوارش و کبد، بیمارستان دکتر شریعتی، دانشگاه علوم پزشکی تهران، تهران، ایران

چکیده

زمینه و هدف:
بیماری سیروز کبدی مهم ترین بیماری کبدی است که سایر بیماری های کبدی به صورت مزمن می‌توانند زمینه ساز بروز آن باشند و خود می‌تواند منشا ابتلا به سرطان کبد شود. این بیماری درمان قطعی ندارد و در مراحل اولیه هیچ گونه نشانه‌ای ندارد و تنها بر اساس آزمایش ها و تصویر برداری ها قابل تشخیص است. در این پژوهش جهت طبقه بندی وضعیت این بیماران بر اساس نشانه‌های آزمایشگاهی از رویکرد داده کاوی استفاده شده است. داده‌کاوی یک علم میان رشته ای است که به کشف دانش پنهان در داده‌ها می‌پردازد.

روش بررسی:
در این مطالعه با استفاده از رویکرد خوشه بندی و الگوریتم k-means وضعیت بیماران را دسته بندی کردیم و برای تعیین کیفیت خوشه بندی و پیدا کردن بهترین تعداد خوشه از شاخص سیلوئت استفاده شد.

یافته‌ها:
داده ها ی ما شامل ۴۱۰ مورد بیماران مبتلا به سیروز کبدی می‌باشد که از بیمارستان دکتر شریعتی جمع آوری شده اند. تعداد مشخصه های مورد بررسی شامل ۲۱ مورد می‌باشد. در این مطالعه وضعیت بیماران مبتلا به سیروز کبدی به دو دسته اصلی تقسیم شد.

نتیجه گیری:
خوشه بندی بر اساس ۲۱ مشخصه انجام شد و میانگین معیار سیلوئت ۴۱ درصد شد. اما با کاهش تعداد مشخصه ها در نهایت با در نظر گرفتن ۷ مشخصه به کیفیت خوشه بندی قابل قبول با متوسط معیار سیلوئت ۶۴ درصد رسیدیم و وضعیت بیماران بر اساس نشانه‌های آزمایشگاهی به دو دسته تقسیم شد که احتمال ابتلابه سرطان در خوشه اول ۲۳ درصد و در خوشه دوم ۱۴ درصد می‌باشد.

کلید واژه: سیروز کبدی، داده کاوی، خوشه بندی، سیلوئت، K-means

گوارش/ دوره ۱۹، شماره۳/ پاییز ۱۳۹۳/ ۱۹۷-۱۹۱

زمینه و هدف:

داده کاوی[1] به عنوان رویکرد کشف دانش در داده ها می‌کاود تا گنج نهان دانش را پیدا کند. داده کاوی در اواخر سال ۲۰۰۹ به عنوان موضوع تحقیقات پزشکی در سال ۲۰۱۰ ثبت شد. امروزه یکی از مهم ترین کاربرد های داده کاوی در زمینه پزشکی است و به پژوهشگران کمک می‌کند که به بینش عمیقی از مجموعه داده های پزشکی دست پیدا کنند.(۱)در سال های اخیر داده کاوی علاوه بر این که توانسته است در زمینه های مختلف پزشکی در فرایند شناخت بیماری، تشخیص بیماری، پیش بینی روند بیماری، انتخاب روش درمان، پیش بینی طول مدت اقامت در بیمارستان، ارتباط بین بیماری ها و ارتباط بین داروها کمک شایانی به ارائه دهندگان خدمات بهداشت و درمان ارائه نماید،(۷-۲) به دلیل افزایش کیفیت خدمات بهداشتی و درمانی باعث کاهش ریسک های ناشی از تشخیص غلط بیماری و یا روش درمان و درنهایت کاهش هزینه و بهبود وضعیت سلامت بیماران شده است.

بر حسب این که در فرایند داده کاوی استنتاج چه نوع دانشی از مجموعه آموزشی مورد نظر است از روش های مختلف داده کاوی می‌توان

1. Data mining

نویسنده مسئول: محمد مهدی سپهری
تهران، بزرگراه جلال آل احمد، پل نصر، دانشگاه تربیت مدرس،
کد پستی ۱۴۱۱۷۱۳۱۱۴
تلفن و نمابر: ۰۲۱-۸۲۸۸۳۳۷۹
پست الکترونیک : mehdi.sepehri@modares.ac.ir
تاریخ دریافت : ۹۳/۲/۱۵
تاریخ اصلاح نهایی : ۹۳/۵/۲۰
تاریخ پذیرش : ۹۳/۵/۲۱

بهره جســت. این روش ها از نظر شیوه یادگیری به دو دسته اصلی تقسیم می‌شوند.(۸)

الف-الگوریتم های یادگیری با نظارت

ب- الگوریتم های یادگیری بدون نظارت

در روش هــای داده کاوی مبتنی بــر الگوریتم های یادگیری با نظارت هدف از داده کاوی مشــخص اســت و در حقیقت تحلیل گر می‌داند که به دنبال چه دانشــی می‌گردد. بر خلاف آن در روش های مبتنی بر یادگیری بدون نظارت، هدف از کشف دانش کاملا تعریف شده نیست.(۹)

خوشــه بندی یک تکنیک داده کاوی است که از رویکرد یادگیری غیر نظارتی برای تحلیل داده ها اســتفاده می‌کند.(۱۰) داده ها بر اســاس اصل حداکثر کردن شباهت داخل گروه ها و حداقل کردن شباهت بین گروه ها، خوشه بندی می‌شــوند. خوشه بندی با پیدا کردن روابط پنهان در داده ها کــه برای پیدا کردن الگوها و پیش بینی وقایع آینده یا تبدیل رکورد داده ها به خوشه های معنی دار مفید هستند، به ارزش پایگاه داده های موجود می‌افزاید.(۱۱)

موضوع مورد بحث در این پژوهش از حوزه بهداشت و سلامت، بیماری سیروز کبدی[1] است. سیروز کبدی یکی از جدی ترین بیماری های کبدی اســت که تقریبا تمام بیماری های کبدی اگر منجر به از بین رفتن ســلول های کبدی شوند می‌توانند به سیروز تبدیل شوند.(۱۲) این بیماری در اثر التهاب مزمن بافت کبد ایجاد می‌شود که به دلایل گوناگونی اتفاق می‌افتد. هر گونه التهابی به صورت کم یا زیاد باعث تخریب بافت کبد می‌شود. در اثر ادامه این فرایند به طور مزمن و مکرر، ســاختار منظم و یک شکل کبد در هم می‌ریزد و به تدریج بافت کبد ســفت می‌شود. فعالیت های احیا کننده تحت سیروز کبدی ادامه می‌یابند ولی سرعت پیشرفت از دست دادن سلول ها بیش از جایگزینی آنها است.(۱۳) علل عمده سیروز کبدی شامل هپاتیت B، هپاتیت C و هپاتیت اتوایمیون[2] اســت. ولی داروها و مواد ســمی،کبد چرب، بیماری ویلســون، تنگی مجاری صفراوی، مشــکلات قلبی و انسداد وریدهای کبدی از علل دیگر این بیماری می باشــند.(۱۲) در مراحل اولیه، بیماری هیچ گونه نشانه‌ای ندارد و تنها با آزمایش های دقیق و بررسی‌های رادیولوژیکی و بافت شناسی قابل تشخیص است.(۱۴) به تدریج که بیماری پیشــرفت می‌کند نشــانه‌ها و عوارض بیماری نیز افزایــش می‌یابد. از مهم ترین نشانه‌های سیروز کبدی آسیت[3] و تجمع مایع در شکم، آنسفالوپاسی کبــدی[4]، واریس مری[5] و خونریزی واریس[6] و یا عوارض دیگری که بر روی سیستم قلب و عروق، ریه ها و کلیه ها ایجاد می‌کند، می‌باشد.(۱۵) درمان بیماری ســیروز کبدی بر اســاس درمان بیماری زمینه ای آن می‌باشــد و درمان قطعی ندارد(۱۶).

1. Liver Cirrhosis
2. Autoimmune hepatitis
3. Ascites
4. Encephalopathy
5. Varices
6. Varical bleeding

سیروز کبدی می‌تواند منجر به ابتلا به سرطان کبد شود که هپاتوسلولار کارســینوما(HCC) رایج ترین نوع آن می‌باشــد. بر اساس گزارش آژانس بین المللی تحقیقات سرطان، HCC در مردان پنجمین سرطان شایع در دنیاســت و در زنان هفتمین سرطان شایع است.(۱۷) این سرطان سومین علت شــایع مرگ و میر در اثر سرطان در دنیا می‌باشد.(۱۸) میزان ابتلای بیماران ســیروتیک به HCC در غرب سالانه بین ۲ تا ۳ درصد و در خاور دور بین ۶ الی ۱۱ درصد می‌باشد.(۱۹)

در این پژوهش به کمک رویکرد خوشــه بندی به دنبال دســته بندی وضعیت بیماران ســیروز کبدی بر اســاس نشــانه‌های آزمایشگاهی آن ها هســتیم. که می‌تواند به درک پزشکان از وضعیت بیماری بیماران بر اساس چند مورد از نشانه‌های آزمایشگاهی و تصمیم گیری آن ها در فرایند درمان تاثیر به سزایی داشته باشد.

روش بررسی:

معرفی داده ها:

کل داده های جمع آوری شده شامل ۴۱۰ رکورد است که از بیماران مبتلا به سیروز مراجعه کننده به بیمارستان دکتر شریعتی جمع آوری شده اســت. داده های جمع آوری شده مربوط به سال های ۸۹ تا ۹۲ می‌باشند و ۵۰ مورد از این بیماران در طی این ۴ ســال به HCC مبتلا شدند. تعداد مشخصه ها ۲۱ مورد است که در جدول ۱ معرفی شده اند.

آماده سازی داده و پیش پردازش:

اغلب به دلیل خطاهای عملیاتی و پیاده سازی سیستم ها، داده های به دســت آمده پر غلط، ناقص و ناسازگارند. بنابراین نیاز است که در ابتدا این داده ها تمیز شوند که این کار با اجرای مراحل پیش پردازش داده ها و در طی مراحل زیرانجام شد:

الف) پاکسازی داده ها:

پاکسازی داده در واقع مرحله کنترل کیفی قبل از تحلیل داده است. در این پژوهش قبل از انجام عملیات پیش پردازش کلیه داده ها از نظر صادق بودن با مقادیر امکان پذیر کنترل شده اند و مقادیر اشتباه اصلاح شده اند. و هم چنین رکوردهایی از داده ها که بسیاری از مشخصه های آن ها خالی بود و یا اعداد اشتباه وارد شده بودند حذف شدند و در این مرحله تعداد کل رکوردهای ما به ۳۶۴ مورد رسید.

پر کردن مقادیر مفقوده:

در ایــن مرحلــه برای پر کردن مقادیر از دســت رفتــه از الگوریتم دســته بندی k نزدیک ترین همســایگی[7] استفاده شــد و بر اساس تعداد ۳ نزدیکترین همســایگی و مد آن ها، مقادیر مفقود شــده پر شدند. علت انتخاب عدد ۳ این بود که انتخاب تعداد کمتر ممکن اســت روش را متاثر

7. K-Nearest Neighbor (KNN)

جدول ۱: مشخصه های مورد بررسی

شماره	۱	۲	۳	۴	۵	۶	۷	۸	۹
مشخصه	Age	Height	Weight	BMI	Platelet	WBC	AST	ALT	Serum Albumin
شماره	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸
مشخصه	Alkph	BiLirubinT	BilirubinD	Creatinine	BUN	INR	FBS	TG	Cholesterol
شماره	۱۹	۲۰	۲۱						
مشخصه	AFP	Na	k						

جدول ۲: تفسیر مقادیر میانگین معیار سیلوئت (۱۶)

میانگین معیار سیلوئت	تفسیر
۱-۰/۷۱	ساختار قوی
۰/۷-۰/۵۱	ساختار منطقی (مناسب)
۰/۵-۰/۲۵	ساختار ضعیف
<۰/۲۵	هیچ ساختار قابل توجهی وجود ندارد

از داده های مغشـوش کرده و مقدار بزرگتر آن هم ممکن است رفتار های نزدیک تر را در نظر نگیرد.(۸)

شناسایی نقاط پرت:

جهت شناسایی نقاط پرت از الگوریتم خوشه بندیK-means استفاده شد. و داده ها به ۴ خوشه تقسیم شدند و ۲۰ داده ی پرت شناسایی و حذف شدند. در این مرحله تعداد داده های ما به ۳۴۴ مورد رسید.

ب) یکپارچه سازی داده:

این عملیات شامل یکپارچه سازی چندین پایگاه داده است. که در این جا داده های مربوط به بیمارسـتان شـریعتی و بیمارستان امام خمینی در یک فایل اکسل مرتب شدند.

ج) تبدیل داده:

در این مرحله فرایند نرمال سازی روی داده ها انجام شد. نرمال سازی تغییر مقیاس داده ها به گونه ای است که آن ها را به یک فاصله ی کوچک و معین نگاشت می کند و باعث می‌شود که داه ها با مقیاس بزرگ نتایج را به سمت خود منحرف نکنند. در این پژوهش با استفاده از روش Min-Max (۸) نرمال سازی داده های کمی که شامل مشخصه های نتایج آزمایشگاهی بود انجام شد و داده ها در بازه عددی بین صفر و یک قرار گرفتند.

مدل سازی:

روش مدل سازی در این پژوهش بر اساس الگوریتم افرازی K-means می‌باشد.

الگوریتم K-means :

پارامتر k (تعداد خوشـه) را به عنوان ورودی گرفته و مجموعه ی n شیء را به k خوشه افراز می‌کند. به طوری که سطح شباهت داخلی خوشه ها بالا بوده و سـطح شـباهت اشیا بین خوشـه ها پایین باشد. شباهت هر خوشه نسبت به متوسط اشیا آن خوشه سنجیده شده که این متوسط مرکز خوشه نیز نامیده می‌شود.(۱۶)

اعتبار سنجی خوشه ها:

شاخص‌های اعتبارسنجی برای سنجش میزان صحت نتایج خوشه‌بندی به منظور مقایسـهٔ بین روش های خوشه‌بندی مختلف یا مقایسه ی نتایج حاصل از یک روش با پارامترهای مختلف مورد استفاده قرار می‌گیرند.

دو معیـارِ پایـهٔ اندازه‌گیری پیشـنهاد شـده برای ارزیابـی و انتخاب خوشه‌های بهینه عبارتند از:

تراکم[1]: داده‌های متعلق به یک خوشـه بایسـتی تا حد ممکن به یکدیگر نزدیک باشـند. معیار رایج برای تعیین میزان تراکم داده‌ها واریانس داده‌ها است.

جدایی[2]: خوشه‌ها خود بایستی به اندازه کافی از یکدیگر جدا باشند.

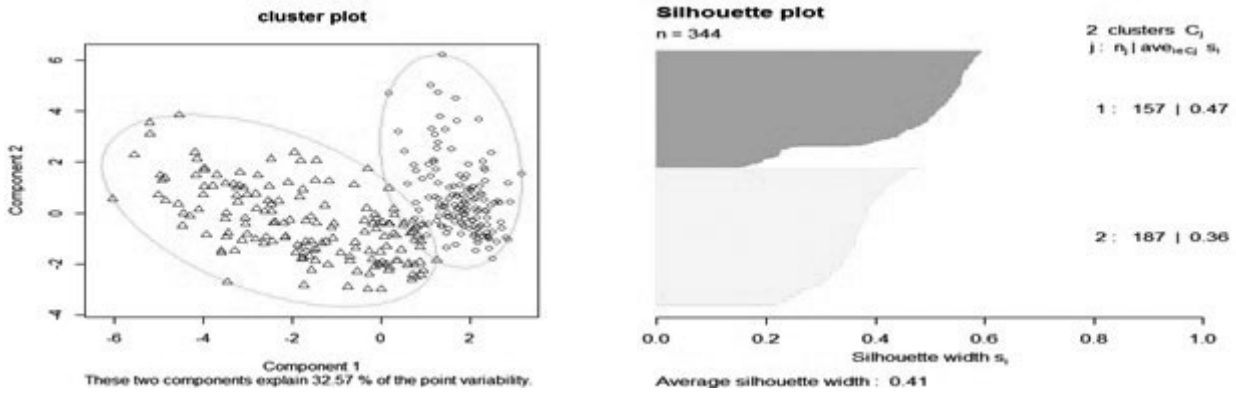
در این پژوهش از شـاخص اعتبار سنجی سیلوئت[3] استفاده شده است که در ادامه مختصرا تشریح می‌شود:

شـاخص سـیلوئت: یکی از معیارهای متداول اعتبارسنجی خوشه بندی اسـت. و دو معیار فواصل درون خوشه ای و برون خوشه ای را همزمان در نظر می‌گیرد.(۱۷) جدول ۲ تفسـیر مقادیر مختلف معیار سیلوئت را نشان می‌دهد:

یافته‌ها:

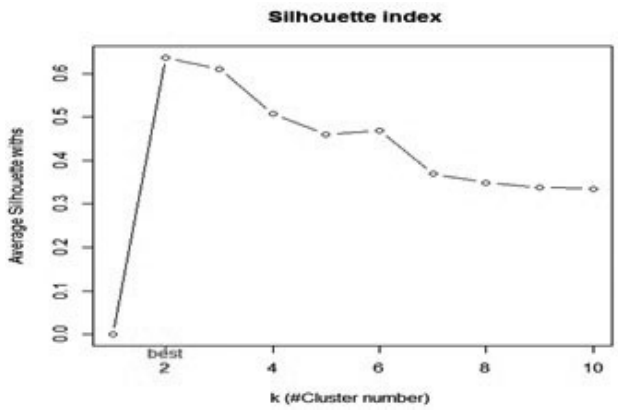
در ابتدا مدل سـازی بر اسـاس ۲۱ مشـخصه جدول ۱-۳ و بر اساس الگوریتم K-means انجام شد . شکل ۱ نتایج با توجه به نمودار سیلوئت در شـکل ۱، در این مرحله متوسـط کیفیت خوشـه مدل خوشه بندی را

1. Compactness
2. Separation
3. Silhouette



ب الف

شکل ۱: (الف) نمودار معیار سیلوئت برای دو خوشه ،(ب) شکل خوشه ها.



شکل ۲: نمودار روند میزان تغییرات معیار سیلوئت برای تعداد خوشه های مختلف.

جدول ۳: مشخصه های مربوط به مدل با ۷ مشخصه

شماره	۱	۲	۳	۴	۵	۶	۷
مشخصه	AST	ALT	Serum Albumin	Creatinine	AFP	Na	K

نشـان می‌دهد. ها ۴۱ درصد اسـت. که بر اسـاس جدول ۲-۳ این خوشه بندی ضعیف است.

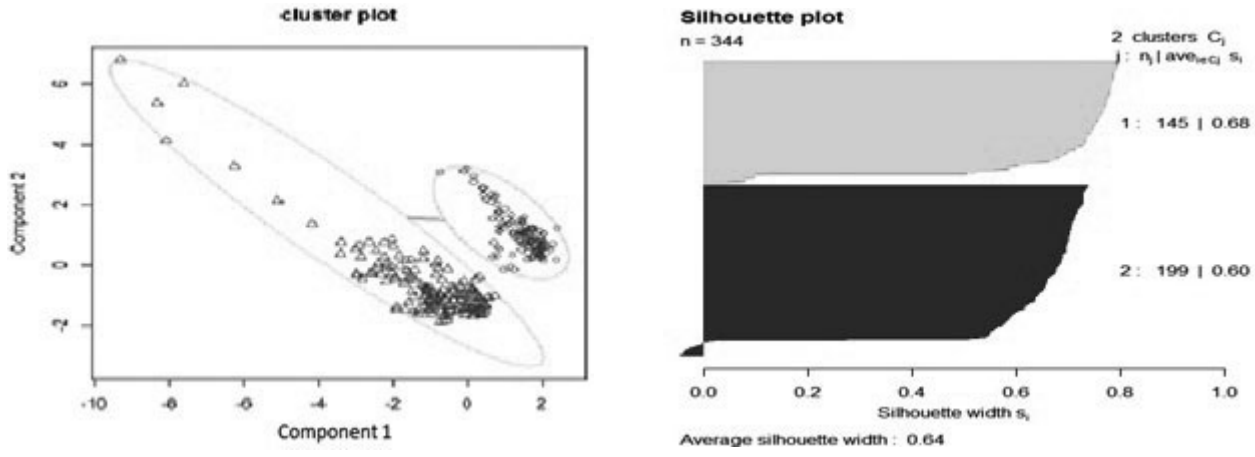
نمودار شـکل ۲ روند تغییرات معیار سـیلوئت با درنظر گرفتن تعداد خوشـه های مختلف را مشـخص می‌کند که بهترین تعداد خوشه در این مرحلـه ۲ می‌باشـد. با کاهش تعداد مشـخصه ها و حذف مشـخصه های ناسـازگار مدل خوشـه بندی بهبـود پیدا کرد و در نهایـت به مدلی با ۷ مشـخصه ی معرفی شده در جدول ۳ رسیدیم. نتایج مدل خوشه بندی در این مرحله در شکل ۳ آورده شده است:

بر اسـاس نمودار سیلوئت در شـکل ۳ برای تعداد دو خوشه ، متوسط کیفیت خوشـه بندی۶۴ درصد اسـت و در این مدل، خوشـه بندی ما در محدوده ی قابل قبول قرار می‌گیرد. شـکل ۴ نمودار روند معیار سیلوئت با درنظر گرفتن تعداد خوشـه های مختلف را نشان می‌دهد. بر اساس شکل ۴ بهترین تعداد خوشه در این مرحله ۲ می‌باشد. بنابراین با در نظر گرفتن نشانه‌های آزمایشگاهی بیماران مبتلا به سیروز کبدی توانستیم وضعیت آن ها را به دو دسته طبقه بندی کنیم. جدول ۲-۴ متوسط مقدار هر مشخصه را در هر خوشه نشان می‌دهد.

براسـاس نتایج فوق وضعیت بیماران به دو دسـته تقسـیم شد که با توجه به جدول ۴ به طور متوسـط خوشه ی۱ وضعیت بدتری را نسبت به خوشه ی ۲ نشان می‌دهد و ۷ مشخصه ی، K ،Na ، AFP ،ALT ،AST و Serum Albumin و Creatinine در بین نشانه‌های آزمایشگاهی در نظر گرفته شده برای تعیین وضعیت بیماران از اهمیت بالاتری برخوردارند.

بحث:

بیماری سیروز کبدی مهم ترین بیماری کبدی است که سایر بیماری های کبدی به صورت مزمن می‌توانند زمینه ساز بروز آن باشند و خود می‌تواند منشا ابتلا به سرطان کبد شود. در این پژوهش جهت بررسی مهم ترین نشانه‌های آزمایشگاهی در تعیین وضعیت بیماران مبتلا به سیروز کبدی و دسته بندی وضعیت آن ها به کمک رویکرد های غیر نظارتی و بر اساس رفتار داده ها، از رویکرد خوشه بندی استفاده کردیم. برای این منظور نتایج آزمایش مربوط به ۴۱۰ بیمار مبتلا به سیروز کبدی در نظر گرفته شد و وضعیت بیماران مبتلا به سیروز کبدی را بر اساس نشانه‌های آزمایشگاهی آنها و رویکرد خوشه بندی که یک رویکرد کاملا غیر نظارتی است تقسیم بندی نمودیم. کلیه نشانه‌های آزمایشگاهی در نظر گرفته شده در این مطالعه بر اساس نتایج آزمایش اولین مراجعه بیماران می‌باشد ولی آنزیم های کبدی TSA و TLA به علت این که در اثر مصرف دارو ها و بیماری های عفونی و بیماری های ایسکمیک و... مقادیرشان تغییر می‌کند،

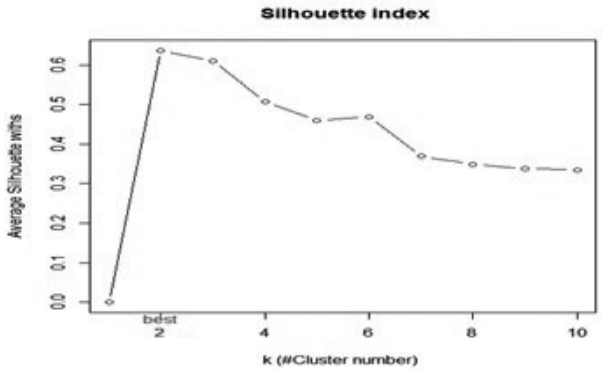


ب الف

شکل ۳: (الف) نمودارمعیار سیلوئت برای دو خوشه، (ب) شکل خوشه ها.

جدول۴: میانگین مقادیر مشخصه ها در خوشه های مختلف

مشخصه	AST		ALT		AFP		Na		K		Serum Albumin		Creatinine	
خوشه	۱	۲	۱	۲	۱	۲	۱	۲	۱	۲	۱	۲	۱	۲
میانگین	۸۷/۴	۶۶/۹	۶۵/۴	۶۰/۱	۳۳/۱	۱۶/۶	۱۳۰	۱۳۸/۸	۴/۵	۴	۳	۳/۵	۱/۲	۱



شکل ۴: نمودار روند میزان تغییرات معیار سیلوئت برای تعداد خوشه های مختلف.

میانگین آن ها در آزمایش های مربوط به هر بیمار در نظر گرفته شد. در رویکرد خوشه بندی اطلاعات گذشته درباره میزان اهمیت نشانه‌ها و داده‌ها در نظر گرفته نمی‌شود و مدل خودش به طور کاملا غیر نظارتی و بر اساس رفتار داده‌ها، نشانه‌های مهم را پیدا می‌کند و دسته بندی را بر اساس آن‌ها انجام می‌دهد. برای انجام خوشه بندی از الگوریتم K-means استفاده شد و بر اساس شاخص اعتبارسنجی سیلوئت کیفیت خوشه بندی و بهترین تعداد خوشه را پیدا کردیم. در ابتدا خوشه بندی بر اساس ۲۱ مشخصه انجام شد که میانگین معیار سیلوئت در این حالت ۴۱ درصد بود و ساختار خوشه بندی ضعیف بود. با حذف مشخصه های ناسازگار کیفیت خوشه بندی بهبود یافت و در نهایت با در نظر گرفتن ۷ مشخصه به کیفیت خوشه بندی قابل قبول با میانگین معیار سیلوئت ۶۴ درصد رسیدیم و وضعیت بیماران بر اساس ۷ مورد از نشانه‌های آزمایشگاهی به دست آمده در این مرحله به دو دسته تقسیم شد. دسته اول با توجه به نشانه‌های در نظر گرفته شده وضعیت بدتری را نشان می‌دهد هم چنین این رویکرد نشان می‌دهد که ۷ مشخصه ی AST، ALT، AFP، Na، K، Serum Albumin و Creatinine در بین نشانه‌های آزمایشگاهی در نظر گرفته شده برای تعیین وضعیت بیماران از اهمیت بالاتری برخوردارند. در نتیجه‌ی مقایسه‌ی میزان ابتلا به سرطان در دو گروه بدست آمده، ملاحظه شد احتمال ابتلا به سرطان در گروه اول ۲۲/۶۹ درصد و در گروه دوم ۱۴ درصد می‌باشد بنابراین ریسک ابتلا به سرطان در گروه اول به میزان تقریبا ۹ درصد بالاتر می‌باشد.بر این اساس به علت این که داده های ما مربوط به رفتار ۴ سال بیماران می‌باشد، احتمال ابتلا به سرطان در گروه اول به طور سالانه حدود ۵/۶ درصد و در گروه دوم حدود ۳/۵ درصد می‌باشد. بر اساس این نتایج می‌توانیم نتیجه گیری کنیم که اگر نشانه‌های آزمایشگاهی بیماری در محدوده گروه اول قرار گیرد احتمال ابتلابه سرطان در این گروه بالاتر می‌باشد و باید درمان های جدی تری برای آن ها در نظر گرفت.

تمرکز این تحقیق بر روی داده کاوی و کاربرد آن در مورد آزمایش های روتین بیماران سیروتیک بوده که مرتبا در مورد این بیماران انجام می‌شود

. گرچه انجام این آزمایش ها تاثیر فوری برای مراقبت از بیماران دارد ولی تاثیر این آزمایش ها در عاقبت بیماران و مراقبت طولانی مدت آنها کاملا مشخص نمی‌باشد. بنابراین در این پژوهش هدف تنها بررسی وضعیت بیماران بر اساس نشانه‌های آزمایشگاهی بوده است و هم چنین کشف غیر نظارتی نشانه‌های مهم در تعیین وضعیت بیماران و باید صحت نتایج آن در مورد بیماران جدید که به درمانگا ه ها مراجعه می‌کنند بررسی شود. این مطالعه، یک مطالعه مقدماتی است که می‌تواند در تحقیقات آتی با در نظر گرفتن نشانه‌های بالینی بیماران توسعه داده شود.

این پژوهش به علت کم بودن تعداد داده ها و تعداد بیماران مبتلا به HCC و هم چنین این که مقادیر آنزیم های AST و ALT در اثر عوامل مختلف بسیار تغییر پذیر می‌باشند و این که نتایج آزمایش بیماران مربوط به آزمایشگاه های مختلف است و می‌تواند خطاهایی را به همراه داشته باشد، ضعف هایی دارد. از نقاط قوت این پژوهش می‌توان به این که اولین پژوهشی است که به کاربرد داده کاوی در بیماری سیروزکبدی پرداخته است و این که در ایران اولین پژوهشی است که به طور متوالی داده های مربوط به چهار سال بیماران سیروتیک را مورد بررسی قرار داده است، اشاره کرد.

REFERENCES

1. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, et al. Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst* 2012;36:2431-48.
2. Anunciaçao O, Gomes BC, Vinga S, Gaspar J, Oliveira AL, Rueff J. A Data Mining Approach for the Detection of High-Risk Breast Cancer Groups. *Adv Bioinformatics* 2010;74:43-51.
3. Gauthier E, Brisson L, Lenca P, Ragusa S. Breast cancer risk score: a data mining approach to improve readability. *International Conference on Data Mining* 2011:15-21.
4. Bhatla N, Jyoti, K. An Analysis of Heart Disease Prediction using Different Data Mining Techniques. *International J Engineering* 2012;1:8.
5. Agrawal A, Misra S, Narayanan R, Polepeddi L, Choudhary A. A lung cancer outcome calculator using ensemble data mining on SEER data. In Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics 2011:5.
6. Kaur AR. A Study of Detection of Lung Cancer Using Data Mining Classification Techniques. *International J* 2013; 3(3).
7. Ahmed K, Jesmin T, Zamilur Rahman M. Early Prevention and Detection of Skin Cancer Risk using Data Mining. *International J Computer Applications* 2013;62:1-6.
8. Han J, Kamber M, Pei J. Data mining: concepts and techniques. *Morgan kaufmann* 2006.
9. Marinakis Y, Marinaki M, Matsatsinis N. A stochastic nature inspired metaheuristic for clustering analysis. *International J Business Intelligence Data Mining* 2008;3:30-44.
10. Wagstaff K, Cardie C, Rogers S, Schrödl S. Constrained k-means clustering with background knowledge. *ICMl* 2001 ;1:577-84.
11. Suh SC, Saffer S, Adla NK. Extraction Of Meaningful Rules In A Medical Database. Seventh International Conference 2008:450-6.
12. Rajeswari P, Reena G. Analysis of Liver Disorder Using Data mining Algorithm. *Global J Computer Science Technology* 2010;10:4.
13. D'Amico G, Garcia-Tsao G, Pagliaro L. Natural history and prognostic indicators of survival in cirrhosis: a systematic review of 118 studies. *J Hepatol* 2006;44:217-31.
14. Singh V, Nagpal S. A Guided clustering Technique for Knowledge Discovery–A Case Study of Liver Disorder Dataset. I*nternational J of Computing Business Research* (1).
15. Asrani SK, Kim WR. Organ allocation for chronic liver disease: model for end-stage liver disease and beyond. *Curr Opin Gastroenterol* 2010;26:209-13.
16. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J computational applied Mm athematics*1987;20:53-65.
17. El-Serag HB. Epidemiology of viral hepatitis and hepatocellular carcinoma. *Gastroenterology* 2012;142:1264-73.
18. Pietrangelo A, Dohil R, Newbury R, Fox L, Bastian J, Aceves S. Reviews Basic Clinc Gastroenterology.
19. Chiaramonte M, Stroffolini T, Vian A, Stazi MA, Floreani A, Lorenzoni U, et al. Rate of incidence of hepatocellular carcinoma in patients with compensated viral cirrhosis. *Cancer* 1999;85:2132-7.