1st International Conference
on new achievements in
Science & Technology
24th August 2017
اولین کنفرانس بین المللی
دستاوردهای نوین در
علوم و تکنولوژی
۲ شهریور ماه ۱۳۹۶ – تهران

کاربرد تکنیک های داده کاوی در تجزیه و تحلیل ویژگی های آکوستیک صدا

مجتبی تلافی داریانی[۱]، سید حسین خبیری[۲]

۱- دانشجوی کارشناسی ارشد، دانشکده مدیریت، دانشگاه تهران، ایران

mojtabatalafi@yahoo.com

۲- دانشجوی کارشناسی ارشد، دانشکده مدیریت، دانشگاه تهران، ایران

hosseinkhabiri@ut.ac.ir

چکیده

تجزیه و تحلیل ویژگی های آکوستیک صدا و گفتار یکی از حوزه هایی است که در سال های اخیر، داده کاوی به آن راه پیدا کرده است. پژوهش حاضر نیز مرتبط با این موضوع است که هدف آن تشخیص جنسیت گوینده با استفاده از ویژگی-های آکوستیک صدای وی می باشد. در این تحقیق مجموعه داده ای شامل ۳۱۶۸ نمونه صدای ضبط شده از سخنرانان زن و مرد انتخاب شد که به واسطه تجزیه و تحلیل آکوستیک، ۲۰ مشخصه به همراه برچسب مورد نظر، استخراج و مهیای اجرای فرآیند داده کاوی گشته است. در نهایت با استفاده از ابزار زبان برنامه‌نویسی پایتون، ۶ تکنیک مختلف شامل: ماشین‌های بردار پشتیبان، رگرسیون لجستیک، جنگل تصادفی، درخت های طبقه بندی و رگرسیونی، بوستینگ تطبیقی و روش K نزدیک ترین همسایه برای ساخت مدل مناسب جهت حل مسأله به کار گرفته شد و دقت مدل ها با یکدیگر مقایسه گردید. آنچه بدست آمد روشن ساخت که تمامی این تکنیک‌ها برای حل مسأله مورد نظر از دقت کافی (بالای ٪۹۰) برخوردار بوده و مدل ساخته شده به وسیله آن ها کارایی لازم جهت کلاسه‌بندی را دارا می باشد. علاوه بر این موضوع، به صورت خاص به بررسی مدل بدست آمده از طریق درخت تصمیم گیری پرداخته و نمونه هایی از قواعد موجود در این مدل را استخراج نمودیم که در نتیجه آن مشخص شد متوسط فرکانس بنیادی اندازه گیری شده در سراسر سیگنال صوتی، جهت بررسی جنسیت صوت، مشخصه کلیدی و ریشه‌ای می باشد به طوری که نقشی محوری در طبقه‌بندی داده ها ایفا می نماید.

واژه های کلیدی: داده کاوی، صدا کاوی، آنالیز گفتار، تحلیل آکوستیک، الگوریتم های طبقه بندی

1st International Conference on new achievements in Science & Technology
24th August 2017
اولین کنفرانس بین المللی دستاوردهای نوین در علوم و تکنولوژی
۲ شهریور ماه ۱۳۹۶ – تهران

Application of data mining techniques in the acoustic analysis of voice

Mojtaba Talafi Daryani[1], Seyyed Hossein Khabiri[2]
1-Department of Management, University of Tehran, Iran
mojtabatalafi@yahoo.com
2-Department of Management, University of Tehran, Iran
hosseinkhabiri@ut.ac.ir

Abstract

One of the areas that data mining technique have been being utilized since last few years is analyzing the acoustic properties of the voice and speech. This study corresponds to this topic and its aim is to distinguish the speaker`s gender by acoustic properties analysis. A dataset including 3168 instances of recorded voices of both female and male speakers was selected and by acoustic analysis, 20 characteristics with their labels are extracted and prepared for data mining. Finally, 6 different techniques including support vector machines (SVM), logistic regression, random forest, classification and regression trees (CART), adaptive boosting (AdaBoost) and K-nearest neighbors' method (KNN) are applied by python programming language and the models' accuracies are compared. Results indicate that all of these techniques and their models are accurate enough (more than 90%) for classifying the data effectively. In addition to that, the study, specifically scrutinizes decision tree model and extracts some existing rules which can be concluded that the average of fundamental frequency measured across acoustic signal plays a significant role in classification and therefore is a key characteristic to gender recognition by voice.

Keywords: *data mining, audio mining, speech analysis, acoustic analysis, classification algorithms*

مقدمه

در سال های اخیر، شاهد این بوده ایم که تکنیک های داده کاوی برای کشف الگو های پنهان در مجموعه داده‌های انبوه و در نتیجه حل مسائل و مشکلات موجود در حوزه علوم مختلف مورد استفاده قرار گرفته است. تشخیص بیماری‌های مختلف در حوزه علوم پزشکی، مدیریت ارتباط با مشتریان در حوزه علوم مدیریت بازاریابی و بازرگانی، پیش بینی قیمت یک سهام یا شاخص در حوزه علوم مالی و اقتصادی، پیش بینی یا تحلیل نتایج انتخابات در حوزه علوم اجتماعی و سیاسی و ... از جمله این مسائل می باشد. با گذشت زمان به تدریج شاهد این اتفاق هستیم که تکنیک‌های داده کاوی وارد زمینه های ناشناخته تری از علم و دانش می شود به طوری که با تکیه بر آن می توان مسائل و مشکلاتی را مرتفع نمود که هرگز تصور آن را هم نمی کردیم. به عنوان نمونه استفاده از تکنیک های داده‌کاوی برای تجزیه و تحلیل صدا و گفتار یکی از این موارد می باشد که پژوهش حاضر نیز در این زمینه صورت پذیرفته است. گفتار، رایج ترین شیوه برقراری ارتباط بین انسان ها می‌باشد. موضوع برقراری ارتباط بین انسان و محیط از طریق صدا و تسلط بر ماشین ها با تکیه بر این واسطه، همواره موضوعی جذاب برای محققان جلوه نموده است. این شاخه از تحقیقات که روزگاری از آن با عنوان یک رؤیا نام برده می شد، در حال حاضر تبدیل به واقعیتی شده است که روز به روز گسترده تر و زوایای پنهان آن آشکار تر می گردد.

این گستردگی و وسعت به جایی رسیده است که در موضوع کلی داده کاوی، شاهد ظهور شاخه هایی همچون داده‌کاوی گفتار[۱]، داده‌کاوی صدا[۲]، کاوش در داده‌های محاورات[۳] و کاوش در داده های شنیداری[۴] هستیم که به صورت کلی از آن ها با عنوان روش های داده کاوی گفتار[۵] نام برده می شود (Senthildevi & Chandra, 2012). در سال‌های اخیر نیز شاهد انجام تحقیقاتی بوده ایم که کاربرد این موضوع را در حوزه های علوم مختلف به اثبات رسانده است. مطالعه صورت گرفته توسط همرلینگ و همکاران[۶] در سال ۲۰۱۶ که به کاربرد داده کاوی صدا جهت ارزیابی آسیب شناسی حنجره پرداخته است نمونه ای از این موارد می باشد.

در این پژوهش قصد داریم به بررسی یکی از مسائل آنالیز گفتار مبنی بر تشخیص جنسیت گوینده با توجه به مشخصات صوتی استخراج شده از گفتار وی بپردازیم. برای این منظور تکنیک های مختلف داده کاوی همچون درخت‌های تصمیم گیری، ماشین های بردار پشتیبان، رگرسیون لجستیک و روش های دیگر را مورد استفاده قرار داده ایم که در قسمت ادبیات تحقیق مفاهیم نظری آن ها را معرفی کرده و در قسمت مواد و روش ها به ذکر جزئیات تکنیک های مورد استفاده پرداخته ایم. امید است که این پژوهش نظر محققان مختلف را به این زمینه ی تحقیقاتی که تا حدی از توجه آن ها مخفی مانده، جلب نموده و کارکرد های پنهان آنالیز گفتار را بیش از پیش آشکار سازد.

ادبیات تحقیق

داده کاوی[۷]:

داده کاوی که کشف دانش در پایگاه داده[۸] نیز نامیده می شود، یک فرآیند تحلیلی است که در رشته های مختلف جهت بررسی روابط معنی دار بین متغیر های موجود در مجموعه داده های بزرگ به کار می رود. تجزیه

---

[1] speech data mining
[2] voice data mining
[3] conversation mining
[4] audio mining
[5] data mining methods of speech
[6] Hemmerling et al.
[7] data mining

و تحلیل جریان انبوه داده‌ها منجر به کشف دانش ارزشمند و مفاهیم نظری می شود که به سازمان ها برای بهبود عملیات و اتخاذ تصمیمات سریع و هوشمند کمک می کند (Jha et al., 2016).

تکنیک های داده کاوی:

الف) طبقه بندی[9]:

رایج ترین تکنیک مورد استفاده در داده کاوی، طبقه بندی می باشد. الگوریتم های طبقه بندی به کاربر اجازه می‌دهد تا یک مجموعه داده پرجمعیت را توسط یک مدل و در قالب طبقه های از پیش تعریف شده کلاسه‌بندی نماید. برخی از این مدل های الگوریتمی عبارت اند از: درخت های تصمیم گیری[10]، جنگل تصادفی[11]، شبکه‌های عصبی[12]، طبقه‌بندی بیزین[13]، ماشین های بردار پشتیبان[14]، الگوریتم K نزدیک ترین همسایه[15]، بوستینگ تطبیقی[16] و طبقه بندی بر مبنای قواعد انجمنی[17] (Maksood & Achuthan, 2016).

ب) خوشه بندی[18]:

خوشه بندی یکی دیگر از تکنیک های داده کاوی می باشد که شامل شناسایی خوشه ها و گروه بندی اشیاء مشابه در هر خوشه می شود. اگر بیان شود که تکنیک طبقه بندی یکی از روش های یادگیری نظارت شده[19] است آنگاه باید اذعان نمود که تکنیک خوشه بندی یکی از روش های یادگیری بدون نظارت[20] می باشد. با وجود اینکه در این بخش محققان بیشتر بر الگوریتم‌های جزء بندی شده[21] مانند K-means تمرکز دارند اما خوشه بندی شامل روش‌های دیگری نیز می شود که عبارت اند از: الگوریتم های خوشه بندی سلسله مراتبی[22] مانند CURE, BIRCH، الگوریتم‌های خوشه‌بندی جدولی[23] مانند STING, WaveCluster، الگوریتم های خوشه‌بندی مبتنی بر مدل[24] مانند COBWEB و الگوریتم های خوشه بندی مبتنی بر چگالی[25] مانند DBSCAN (Maksood & Achuthan, 2016).

ج) رگرسیون[26]:

رگرسیون تکنیکی است که برای مدل سازی پیش بینی کننده[27] مورد استفاده قرار می گیرد. هدف تحلیل رگرسیون تعیین بهترین مدلی است که نحوه ارتباط یک متغیر را با یک یا چند متغیر دیگر تعیین می کند. از آنجا که در دنیای واقعی، پیش بینی نیازمند ادغام جنبه های مختلف و پیچیده مجموعه داده می باشد بنابراین جهت تکمیل آن از ترکیب مدل های مختلف استفاده می شود. از جمله این الگوریتم های ترکیبی می توان به

---

[8] Knowledge Discovery in Database (KDD)
[9] classification
[10] decision trees
[11] random forest
[12] neural networks
[13] Bayesian classification
[14] Support Vector Machines (SVM)
[15] K-Nearest Neighbor (KNN)
[16] adaboost
[17] classification based in association
[18] clustering
[19] supervised learning method
[20] unsupervised learning method
[21] partitioned algorithms
[22] hierarchical algorithms
[23] grid-based algorithms
[24] model-based algorithms
[25] density-based algorithms
[26] regression
[27] predictive modeling

درخت‌های طبقه بندی و رگرسیونی[۲۸] اشاره نمود. روش های مختلف رگرسیون که به کار گرفته می شوند عبارت اند از: رگرسیون لجستیک[۲۹]، رگرسیون خطی[۳۰]، رگرسیون خطی چند متغیره[۳۱]، رگرسیون غیر خطی[۳۲] و رگرسیون غیر خطی چند متغیره[۳۳] (Maksood & Achuthan, 2016).

در جدول ۱ می توان مفاهیم نظری مربوط به تکنیک های داده کاوی که در این پژوهش مورد استفاده قرار گرفتند را ملاحظه نمود.

جدول ۱: مفاهیم نظری تکنیک های داده کاوی مورد استفاده

ردیف	روش	توضیح
۱	ماشین های بردار پشتیبان	این روش به وسیله ساخت یک طرح فوق العاده در یک فضای چند بعدی که نمونه‌ها با کلاس ها و برچسب های متفاوت را جداسازی می کند، عمل طبقه بندی را انجام می دهد. این تکنیک از عملیات و وظایف مربوط به رگرسیون و طبقه‌بندی پشتیبانی کرده و علاوه بر این توانایی ارزیابی متغیر های پیوسته و طبقه ای را دارا می‌باشد. این روش بر مبنای مفهوم طرح های تصمیم گیری که حاوی مرز های تصمیم گیری می باشند استوار گشته است (Fatima & Ikbal Khan, 2016).
۲	رگرسیون لجستیک	در این روش جهت مناسب سازی مدل های خطی می توان مدلی را بر داده ها منطبق نمود که تخمین دقیقی از احتمال طبقه بدست می دهد. رگرسیون لجستیک از مدل های خطی برای طبقه بندی و از رگرسیون خطی برای تخمین مقدار عددی هدف استفاده می کند (Provost & Fawcett, 2013).
۳	جنگل تصادفی	جنگل تصادفی یکی از روش های زیر مجموعه درخت های تصمیم گیری است که برای حل مسأله تعداد زیادی درخت تصادفی روی زیر مجموعه هایی از مجموعه داده تولید می نماید و به واسطه میانگین گیری باعث بهبود دقت نتایج می‌شود (Biau & Scornet, 2016).
۴	درخت های طبقه بندی و رگرسیونی	این روش یکی از الگوریتم های درخت تصمیم‌گیری می باشد که از درخت های طبقه بندی جهت کلاسه بندی متغیرهای وابسته و از درخت های رگرسیونی جهت پیش‌بینی متغیر های پاسخ استفاده می‌کند (Maksood & Achuthan, 2016).
۵	بوستینگ تطبیقی	بوستینگ روشی است که بر مبنای دستیابی به یک قانون بسیار دقیق جهت پیش‌بینی از طریق ترکیب کردن تعداد زیادی قوانین ضعیف و نادقیق استوار است. بوستینگ تطبیقی اولین الگوریتم کاربردی بوستینگ و یکی از پرکاربردترین آن ها در مسائل مختلف می‌باشد (Schapire, 2013).
۶	K نزدیک ترین همسایه	این روش بر مبنای یادگیری توسط نمونه های آموزش استوار است. هر نمونه، نماینده یک نقطه در فضای n بعدی می باشد. همه نمونه های آموزش در یک الگوی n بعدی فضایی ذخیره می شود. زمانی که یک نمونه ناشناخته داده می شود، طبقه بند k نزدیک ترین همسایه، به دنبال k نمونه آموزش که به نمونه ناشناخته نزدیک ترین هستند، الگوی فضایی را جستجو می کند. نزدیکی بر اساس فاصله اقلیدسی[۳۴] تعریف می شود. پس از یافتن این k داده مشابه با نمونه آزمایشی، با رأی اکثریت برچسب نمونه ناشناخته انتخاب می شود. همچنین با تخصیص وزن به هر یک از مشخصه‌ها می توان درصد مشارکت آن ها در محاسبه تشابه را تغییر داد (Gorade et al., 2017).

[28] Classification and Regression Trees (CART)
[29] logistic regression
[30] linear regression
[31] multivariate linear regression
[32] nonlinear regression
[33] multivariate nonlinear regression
[34] Euclidean distance

مراحل داده کاوی:

لازم است به این مطلب توجه داشته باشیم که اجرای تکنیک های داده کاوی تنها یک مرحله از سلسله مراحل فرآیند کشف دانش در پایگاه داده می باشد و علاوه بر آن مراحلی وجود دارد که اهمیت دادن به آن ها ضروری به نظر می‌رسد.

در شکل ۱ می توان مراحل کشف دانش در پایگاه داده را ملاحظه نمود. این مراحل عبارت اند از:

الف) انتخاب داده[۳۵]:

این مرحله شامل مطالعه درباره دامنه کاربرد و انتخاب مجموعه داده می باشد. مطالعه دامنه کاربرد در نظر دارد تا از طریق درک یک مسأله کسب و کار اهداف پروژه را مشخص سازد. در این مرحله ضروری است تا حداقل اندازه، مشخصه های مورد نیاز و بازه زمانی مناسب برای مجموعه داده تشخیص داده شود (Sharma & Mittal, 2016).

ب) آماده سازی داده[۳۶]:

این مرحله شامل عملیاتی نظیر پاکسازی داده[۳۷] از طریق حذف داده های پرت، تصمیم گیری در مورد داده‌های مفقود و ... می شود. همچنین در این مرحله می توان به مسائل مربوط به مدیریت پایگاه داده مانند نوع داده، الگوی مقادیر مفقود و ... پرداخت (Sharma & Mittal, 2016).

ج) تبدیل داده[۳۸]:

این مرحله شامل پردازش داده جهت تبدیل آن به قالب مناسب برای اعمال الگوریتم های داده کاوی می باشد. پردازش‌های مرسومی که در این مرحله می توان به آن ها اشاره کرد عبارت اند از: انتخاب مشخصه[۳۹]، نرمال-سازی داده[۴۰]، تجمیع داده[۴۱] و گسسته سازی داده[۴۲]. برای نرمال سازی داده، مقدار میانگین، از هر مقدار کاسته شده و سپس حاصل بر انحراف معیار تقسیم می شود. برخی الگوریتم ها تنها با داده های کمی یا داده های کیفی سازگاری دارند لذا لازم است تا گاهی اوقات نوع داده را تغییر دهیم (Sharma & Mittal, 2016).

د) داده کاوی:

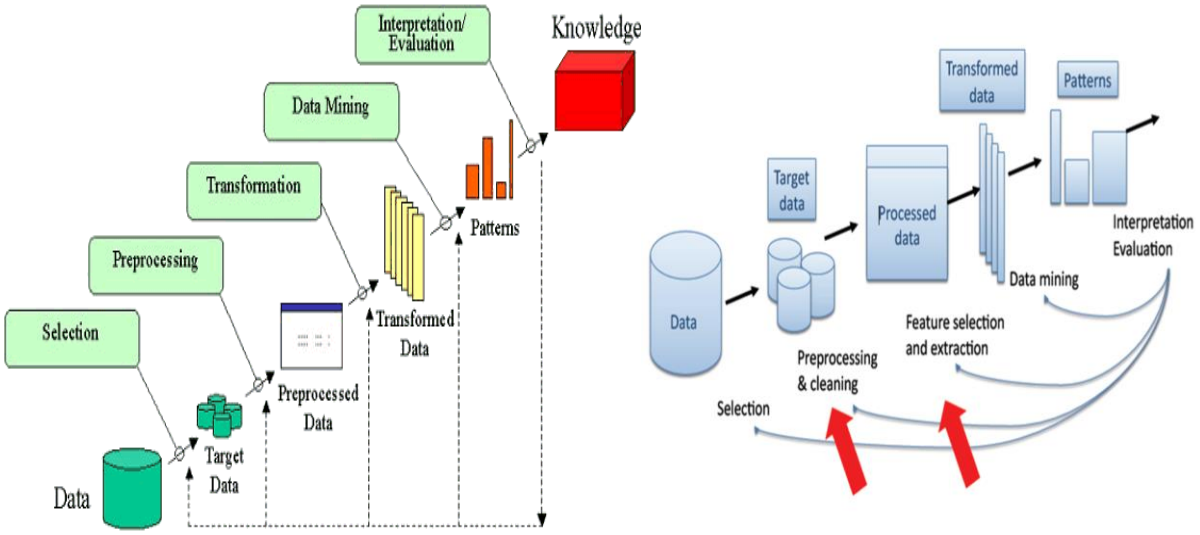
این مرحله شامل کشف الگو های موجود در مجموعه داده ای که در مراحل قبل آماده نمودیم، می شود. در این مرحله الگوریتم های مختلفی مورد ارزیابی قرار می گیرد تا بهترین روش جهت دستیابی به منظور خاصی انتخاب گردد (Sharma & Mittal, 2016).

هـ) تفسیر و ارزیابی نتایج:

این مرحله شامل تفسیر الگو های کشف شده و ارزیابی کاربرد و اهمیت آن ها با توجه به دامنه کاربرد می‌شود. در این مرحله به عنوان نمونه می توان چنین نتیجه ای را برداشت نمود که برخی از مشخصه های انتخاب شده را می توان نادیده گرفت چرا که دخالتی در نتایج بدست آمده و تحلیل صورت گرفته نداشتند بنابراین می‌توان فرآیند را بعد از اصلاح مجموعه داده دوباره تکرار نمود (Sharma & Mittal, 2016).

---

[35] data selection
[36] data pre-processing
[37] data cleaning
[38] data transformation
[39] feature selection
[40] data normalization
[41] data aggregation
[42] data discretization



شکل ۱: مراحل کشف دانش در پایگاه داده

پیشینه تحقیق

در زمینه آنالیز گفتار و تشخیص صدا با استفاده از تکنیک های داده کاوی تحقیقات بسیار انگشت شماری صورت گرفته است. شاید بتوان اذعان نمود که در دسترس نبودن مجموعه داده های کافی در این حوزه از یک طرف و عدم آشنایی محققان با موضوع آنالیز گفتار و کارکرد های داده کاوی در آن از طرف دیگر باعث فقر تحقیقات در این موضوع شده است.

یکی از تحقیقاتی که تا حدی در ارتباط با پژوهش حاضر می باشد، تحقیق صورت گرفته توسط بویوک ییلماز و سیبیکدیکن[۴۳] در سال ۲۰۱۶ است. این تحقیق بر روی مجموعه داده پژوهش حاضر صورت گرفته است به طوری که با استفاده از ابزار زبان برنامه نویسی پایتون به تکمیل و ارائه الگوریتم شبکه های عصبی پرسپترون چند لایه[۴۴] پرداخته است. این روش یک مدل شبکه عصبی مصنوعی پیشخور می باشد که مجموعه داده های ورودی را به مجموعه خروجی های مناسب متناظر می کند. همچنین این روش شامل چند لایه از گره ها در یک گراف جهت دار است به طوری که هر لایه کاملاً به لایه بعدی متصل می باشد. محققان در این مطالعه در نهایت توانستند با دقت ۹۶/۷۴ ٪ به کلاسه بندی مجموعه داده مورد آزمایش بپردازند.

مواد و روش ها

انتخاب داده:

در این پژوهش مجموعه داده ای با عنوان شناسایی جنسیت از طریق صوت[۴۵] از بین مجموعه داده های سایت کَگِل[۴۶] انتخاب شده است. این مجموعه داده به منظور شناسایی صدای زن یا مرد بر مبنای ویژگی های صوتی[۴۷] صدای آن ها تهیه شده است. مجموعه داده شامل ۳۱۶۸ نمونه صدای ضبط شده از سخنرانان زن و مرد می‌باشد.

---

[43] Buyukyilmaz & Cibikdiken
[44] Multilayer Perceptron (MLP)
[45] gender recognition by voice
[46] www.kaggle.com
[47] acoustic properties of the voice and speech

نمونه های صوتی به واسطه ی فرآیند تجزیه و تحلیل صوتی (تحلیل آکوستیک)[48] آماده سازی شده و در نهایت ۲۰ مشخصه به همراه برچسب مورد نظر استخراج شده است. این ویژگی های صوتی که در فایل مجموعه داده ذکر شده اند در جدول ۲ قابل ملاحظه می‌باشد.

جدول ۲: مشخصه های مجموعه داده

ردیف	مشخصه	توضیح	
۱	meanfreq	فرکانس متوسط (kHz)	mean frequency (in kHz)
۲	sd	انحراف معیار فرکانس	standard deviation of frequency
۳	median	فرکانس میانه (kHz)	median frequency (in kHz)
۴	Q25	چارک اول (kHz)	first quantile (in kHz)
۵	Q75	چارک سوم (kHz)	third quantile (in kHz)
۶	IQR	دامنه میان چارکی (kHz)	interquartile range (in kHz)
۷	skew	چولگی	skewness
۸	kurt	کشیدگی	kurtosis
۹	sp.ent	انتروپی طیفی	spectral entropy
۱۰	sfm	صافی طیفی	spectral flatness
۱۱	mode	فرکانس مد	mode frequency
۱۲	centroid	مرکز فرکانس	frequency centroid
۱۳	meanfun	متوسط فرکانس بنیادی اندازه گیری شده در سراسر سیگنال صوتی	average of fundamental frequency measured across acoustic signal
۱۴	minfun	حداقل فرکانس بنیادی اندازه گیری شده در سراسر سیگنال صوتی	minimum fundamental frequency measured across acoustic signal
۱۵	maxfun	حداکثر فرکانس بنیادی اندازه گیری شده در سراسر سیگنال صوتی	maximum fundamental frequency measured across acoustic signal
۱۶	meandom	متوسط فرکانس غالب اندازه گیری شده در سراسر سیگنال صوتی	average of dominant frequency measured across acoustic signal
۱۷	mindom	حداقل فرکانس غالب اندازه گیری شده در سراسر سیگنال صوتی	minimum of dominant frequency measured across acoustic signal
۱۸	maxdom	حداکثر فرکانس غالب اندازه گیری شده در سراسر سیگنال صوتی	maximum of dominant frequency measured across acoustic signal
۱۹	dfrange	دامنه فرکانس غالب اندازه گیری شده در سراسر سیگنال صوتی	range of dominant frequency measured across acoustic signal
۲۰	modindx	شاخص مدولاسیون	modulation index
۲۱	label	برچسب (مرد یا زن)	male or female

ابزار داده کاوی:

[48] acoustic analysis

در پژوهش حاضر کلیه ی پردازش ها، روش ها و تکنیک هایی که جهت حل مسأله مورد نظر به کار گرفته شده و در ادامه توضیح داده می شود، با استفاده از ابزار زبان برنامه نویسی پایتون[۴۹] و کتابخانه های کاربردی آن در محیط ژوپیتر نوت بوک[۵۰] اجرا شده است.

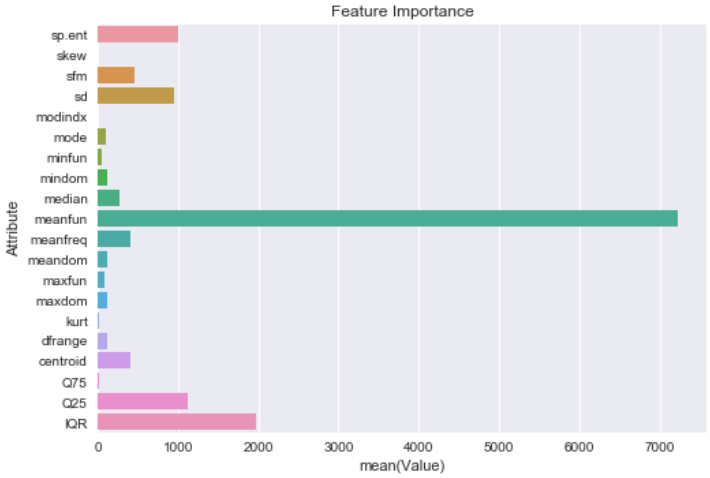
آماده سازی و تبدیل داده:

مرحله پیش پردازش داده معمولاً مورد غفلت واقع می شود اما مرحله ای مهم و مورد نیاز برای اجرای تکنیک-های داده کاوی می باشد. در این مرحله می توان از هر روشی که داده های خام را جهت پردازش های بعدی آماده سازد و آن را جهت استفاده ی آسان و مؤثر بازآرایی کند، استفاده نمود (Bharat et al., 2016). از آنجایی که داده های موجود در دنیای واقعی ممکن است کیفیت لازم برای شروع داده کاوی را نداشته باشند لذا اجرای مرحله آماده سازی و تبدیل داده ضرورت پیدا می کند.

به دلیل اینکه مجموعه داده حاضر، عاری از وجود داده های تکراری یا مقادیر از دست رفته می باشد لذا در این مرحله نیازی به انجام عملیاتی نظیر حذف داده های تکراری، تصمیم گیری در مورد داده های مفقود و ... وجود نداشت. یکی از اهداف مهمی که در این مرحله به دنبال آن بودیم افزایش درک خود از مجموعه داده و مشخصه های موجود می باشد چرا که این مهم، در بهبود کیفیت فرآیند کشف دانش نقش به سزایی ایفا می‌نماید. در همین راستا سعی شده است تا از روش های مختلف اطلاعات خود در مورد داده ها را افزایش دهیم. به عنوان نمونه می توان به بررسی نوع[۵۱] و قالب داده[۵۲]، بررسی فاکتور های آماری داده های هر مشخصه (تعداد، میانگین، انحراف معیار، مقدار بیشینه و کمینه، میانه، چارک اول و سوم)، بررسی همبستگی[۵۳] بین مشخصه ها و روش های بصری سازی داده[۵۴] (نمودار های پراکندگی[۵۵]، نمودار های جعبه ای[۵۶]، نمودار های بافت نگار[۵۷] و ...) اشاره نمود که برای این منظور از آن ها استفاده نمودیم. همچنین در این مرحله برای استفاده از تکنیک های داده-کاوی و بهبود کیفیت نتایج حاصل از آن پردازش‌های نرمال سازی داده و انتخاب مشخصه را اجرا نمودیم. برای انجام فرآیند انتخاب مشخصه مؤثر نیز به بررسی ارزش و اهمیت هر مشخصه پرداخته و در نتیجه آن با انتخاب مشخصه های ارزشمند یک دسته مشخصه جدید برای اجرای تکنیک های داده کاوی در مرحله بعد ایجاد نمودیم. در شکل ۲ می توان میزان اهمیت هر مشخصه را ملاحظه نمود.

ذکر این نکته نیز لازم است که برای اجرای تکنیک های داده کاوی در مرحله بعد، قسمتی از داده ها (۲۰ ٪) جدا شده و در ساخت مدل شرکت داده نشدند چرا که از آن ها بتوان در مرحله آزمایش مدل استفاده نمود.

[49] python programming language
[50] Jupyter Notebook
[51] data type
[52] data frame
[53] correlation
[54] data visualization
[55] scatter plots
[56] box plots
[57] histograms



شکل ۲: ارزش و اهمیت مشخصه ها

داده کاوی:

پس از آن که مجموعه داده جهت اجرای تکنیک های داده کاوی و فرآیند کشف دانش از آن، آماده گردید و درک لازم از مشخصه های موجود و روابط بین آن ها حاصل گشت، در این مرحله تکنیک های مختلف داده‌کاوی جهت مقایسه نتایج آن ها برای حل مسأله مورد نظر در این پژوهش مورد استفاده قرار گرفت. در تحقیق حاضر از تکنیک‌های ماشین های بردار پشتیبان، رگرسیون لجستیک، جنگل تصادفی، درخت های طبقه بندی و رگرسیونی، بوستینگ تطبیقی و K نزدیک ترین همسایه استفاده شده است که نتایج و مقایسه بین آن ها در قسمت یافته ها قابل ملاحظه می‌باشد. همچنین در اجرای هر یک از این تکنیک ها از ترفند های مختلف جهت دستیابی به نتایج بهتر استفاده شده‌است که از جمله آن ها می توان به اعتبار سنجی متقابل[۵۸] و جستجوی حریصانه[۵۹] اشاره نمود. در روش اعتبارسنجی متقابل مجموعه داده به چند بخش تقسیم می شود (معمولاً ۵ یا ۱۰ قسمت) سپس مراحل ساخت و آزمایش مدل به تعداد بخش ها تکرار می شود. در هر بار تکرار یک قسمت به عنوان داده آزمایش[۶۰] انتخاب شده و سایر قسمت ها به عنوان داده آموزش[۶۱] برای ساخت مدل با یکدیگر ترکیب می شوند. در نهایت برای سنجش اعتبار مدل، میانگین دقت مراحل تکرار شده و واریانس آن مد نظر قرار می-گیرد (Provost & Fawcett, 2013). جستجوی حریصانه نیز ترفندی است که با استفاده از آن می توان در برخی تکنیک ها (روش هایی که پارامتر آن ها قابل تنظیم باشد)، پارامتر بهینه را بدست آورد تا بهترین نتیجه حاصل گردد. به عنوان مثال می توان به این مورد اشاره نمود که با استفاده از روش جستجوی حریصانه می توان بهترین عمق ممکن برای درخت تصمیم گیری را بدست آورد تا دقت مد نظر حاصل گردد. در جدول ۳ تکنیک‌های داده کاوی که به کار گرفته شده به علاوه روش های تکمیلی که برای مناسب‌سازی هر تکنیک و بهبود دقت نتایج آن مورد استفاده قرار گرفته است، قابل ملاحظه می باشد.

[58] cross validation
[59] grid search
[60] testing data
[61] training data

جدول ۳: تکنیک ها و روش های داده کاوی مورد استفاده

ردیف	تکنیک های داده کاوی	روش های تکمیلی
۱	ماشین های بردار پشتیبان	انتخاب تابع کرنل[۶۲]، اعتبارسنجی متقابل، جستجوی حریصانه
۲	رگرسیون لجستیک	انتخاب مشخصه، اعتبارسنجی متقابل، جستجوی حریصانه
۳	جنگل تصادفی	انتخاب مشخصه، اعتبارسنجی متقابل، جستجوی حریصانه
۴	درخت های طبقه بندی و رگرسیونی	انتخاب مشخصه، اعتبارسنجی متقابل، جستجوی حریصانه
۵	بوستینگ تطبیقی	انتخاب مشخصه، اعتبارسنجی متقابل، جستجوی حریصانه
۶	K نزدیک ترین همسایه	انتخاب مشخصه، اعتبارسنجی متقابل

یافته ها

پس از آن که تکنیک های مختلف داده کاوی برای حل مسأله مورد نظر به کار گرفته شد، دقت[۶۳] هر روش و همچنین سطح زیر منحنی مشخصه عملیاتی دریافت کننده[۶۴] جهت بررسی نتایج و مقایسه آن ها با یکدیگر محاسبه و گردآوری شده است. منحنی مشخصه عملیاتی دریافت کننده یک معیار سنجش میزان کارایی در مسائل طبقه بندی است. هر چه مقدار سطح زیر این منحنی ها بیشتر باشد کارایی نهایی مدل مطلوب تر ارزیابی می شود. در جدول ۴ می توان دقت روش های مورد استفاده را با یکدیگر مقایسه نمود.

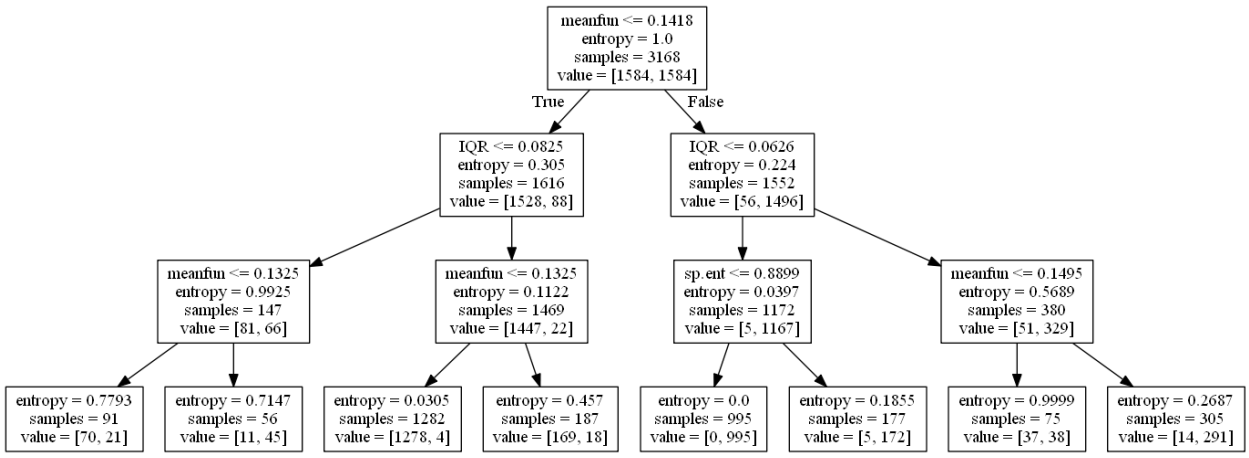
جدول ۴: دقت و سطح زیر نمودار ROC تکنیک ها و روش های داده کاوی مورد استفاده

ردیف	تکنیک های داده کاوی	دقت	سطح زیر نمودار ROC
۱	ماشین های بردار پشتیبان (تابع کرنل خطی)	0.977918	0.977970
۲	ماشین های بردار پشتیبان (تابع کرنل خطی) + اعتبارسنجی متقابل	0.971577	0.977970
۳	ماشین های بردار پشتیبان (تابع کرنل RBF) + اعتبارسنجی متقابل	0.971552	0.976362
۴	ماشین های بردار پشتیبان (تابع کرنل خطی) + اعتبارسنجی متقابل + جستجوی حریصانه	0.975138	0.976362
۵	ماشین های بردار پشتیبان (تابع کرنل RBF) + اعتبارسنجی متقابل + جستجوی حریصانه	0.981452	0.977970
۶	رگرسیون لجستیک	0.970032	0.969991
۷	رگرسیون لجستیک + اعتبارسنجی متقابل	0.968377	0.969991
۸	رگرسیون لجستیک + انتخاب مشخصه	0.963722	0.963680
۹	رگرسیون لجستیک + اعتبارسنجی متقابل + انتخاب مشخصه	0.968402	0.963680
۱۰	رگرسیون لجستیک + اعتبارسنجی متقابل + جستجوی حریصانه	0.972770	0.971539
۱۱	رگرسیون لجستیک + اعتبارسنجی متقابل + جستجوی حریصانه + انتخاب مشخصه	0.973560	0.963680
۱۲	جنگل تصادفی	0.971609	0.971599
۱۳	جنگل تصادفی + اعتبارسنجی متقابل	0.958906	0.971599
۱۴	جنگل تصادفی + انتخاب مشخصه	0.973186	0.973206
۱۵	جنگل تصادفی + اعتبارسنجی متقابل + انتخاب مشخصه	0.973076	0.973206

[62] kernel function
[63] accuracy
[64] Receiver Operating Characteristic (ROC) curve

۱۶	جنگل تصادفی + اعتبارسنجی متقابل + جستجوی حریصانه	0.981452	0.973147
۱۷	جنگل تصادفی + اعتبارسنجی متقابل + جستجوی حریصانه + انتخاب مشخصه	0.979874	0.974993
۱۸	درخت های طبقه بندی و رگرسیونی	0.971609	0.971778
۱۹	درخت های طبقه بندی و رگرسیونی + اعتبارسنجی متقابل	0.957463	0.971778
۲۰	درخت های طبقه بندی و رگرسیونی + اعتبارسنجی متقابل + انتخاب مشخصه	0.954264	0.968383
۲۱	درخت های طبقه بندی و رگرسیونی + اعتبارسنجی متقابل + جستجوی حریصانه	0.970008	0.960464
۲۲	درخت های طبقه بندی و رگرسیونی + اعتبارسنجی متقابل + جستجوی حریصانه + انتخاب مشخصه	0.974349	0.962191
۲۳	بوستینگ تطبیقی	0.974763	0.974814
۲۴	بوستینگ تطبیقی + اعتبارسنجی متقابل	0.968500	0.974814
۲۵	بوستینگ تطبیقی + اعتبارسنجی متقابل + انتخاب مشخصه	0.971625	0.969991
۲۶	بوستینگ تطبیقی + اعتبارسنجی متقابل + جستجوی حریصانه	0.980268	0.976362
۲۷	بوستینگ تطبیقی + اعتبارسنجی متقابل + جستجوی حریصانه + انتخاب مشخصه	0.972376	0.968443
۲۸	K نزدیک ترین همسایه	0.965300	0.965168
۲۹	K نزدیک ترین همسایه + اعتبارسنجی متقابل	0.946303	0.965168
۳۰	K نزدیک ترین همسایه + انتخاب مشخصه	0.981073	0.981066
۳۱	K نزدیک ترین همسایه + اعتبارسنجی متقابل + انتخاب مشخصه	0.966790	0.981066

همان طور که در جدول ۴ قابل ملاحظه می باشد، تمامی تکنیک های داده کاوی استفاده شده برای حل مسأله مورد نظر در پژوهش حاضر از دقت کافی برای ساخت مدل طبقه بندی برخوردار می باشد و از آن جایی که میزان دقت هر روش و سطح زیر منحنی مشخصه عملیاتی دریافت کننده آن ها بسیار نزدیک به هم می باشد (دقت بالای ۹۰ ٪) مقایسه بین تکنیک ها در این پژوهش نتیجه ی خاص و منحصر به فردی بدست نمی دهد. گذشته از این مطلب یکی از یافته هایی که می تواند حائز اهمیت قرار گیرد، توجه به درخت تصمیم گیری بدست آمده و قواعد استخراج شده از آن می باشد که به عنوان یک مدل بدست آمده جهت حل مسأله ی مد نظر که از نوع مسائل طبقه بندی بوده، ارزشمند می باشد. در شکل ۳ می توان درخت تصمیم گیری بدست آمده را مشاهده نمود. همچنین لازم به ذکر است که این درخت تصمیم گیری از نوع درخت های طبقه بندی و رگرسیونی بوده که برای رسم آن از روش انتروپی[۶۵] استفاده شده است.



شکل ۳: درخت تصمیم گیری طبقه بندی و رگرسیونی

[65] entropy

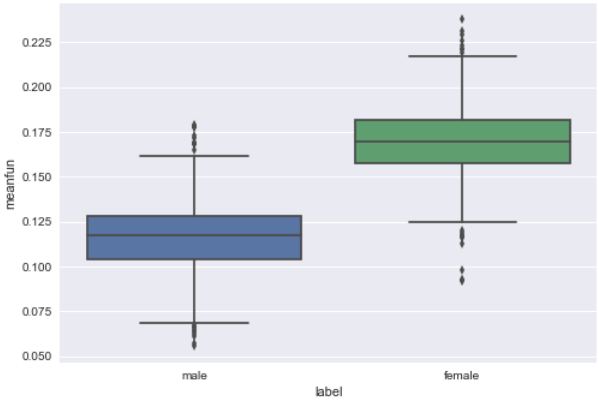
همان طور که در شکل ۳ قابل ملاحظه می باشد، درخت تصمیم گیری حاوی قواعد اگر - آنگاه ارزشمندی است که در حیطه کاربرد می توان آن ها را جهت کلاسه بندی داده ها و تشخیص جنسیت صوت مورد استفاده قرار داد. به عنوان نمونه می توان به این قواعد اشاره نمود:

اگر meanfun <= 0.1418 و IQR >= 0.0825 آنگاه با احتمال 0.985 صدا مربوط به یک مرد می باشد.

اگر meanfun >= 0.1418 و IQR <= 0.0626 آنگاه با احتمال 0.996 صدا مربوط به یک زن می باشد.

بحث و نتیجه گیری

در این پژوهش به دنبال این بودیم که کاربرد داده کاوی در تجزیه و تحلیل آکوستیک صدا را نشان داده و در همین راستا به نتایج قابل توجهی نیز دست پیدا نمودیم. چنانچه گذشت، می توان از تکنیک های مختلف داده-کاوی برای تشخیص جنسیت با استفاده از صدا استفاده نمود به طوری که مدل های حاصل از این تکنیک ها دارای دقت کافی برای طبقه بندی داده ها و برچسب گذاری آن ها می باشد. به عنوان نمونه درخت تصمیم‌گیری حاصل از مدل سازی داده ها را مورد مطالعه قرار دادیم و متوجه شدیم که برای تشخیص جنسیت با توجه به صدا، متوسط فرکانس بنیادی اندازه گیری شده در سراسر سیگنال صوتی، بیشتر از سایر مشخصه ها حائز اهمیت می باشد به طوری که در درخت تصمیم گیری، این مشخصه به عنوان گره اصلی (گره ریشه) برای کلاسه بندی داده ها حاصل شده است. این نتیجه با توجه به پژوهش های صورت گرفته در این حوزه قابل ارزیابی می باشد به طوری که به عنوان نمونه پژوهش انجام شده در سال ۲۰۱۴ توسط پون و ان جی[۶۶] این حقیقت را تصدیق نمود که فرکانس بنیادی، راهنمای اصلی جهت شناسایی صدای مردان و زنان می باشد. این مطلب که فرکانس بنیادی در طبقه بندی داده ها نقش کلیدی ایفا می کند، در شکل ۴ نیز که یک نمودار جعبه ای می باشد قابل ملاحظه است.

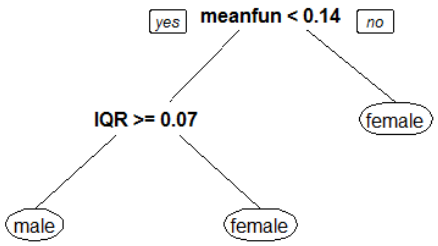


شکل ۴: نمودار جعبه ای فرکانس بنیادی - برچسب

یکی از نکاتی که باید به آن توجه داشته باشیم این است که طبق نتایج بدست آمده از تحقیقات صورت گرفته، فرکانس بنیادی صدای مردان و زنان تفاوت قابل توجهی با یکدیگر دارند و به همین خاطر است که در این مسأله فرکانس بنیادی نقش محوری را جهت کلاسه بندی داده ها بر عهده دارد. بازه فرکانس بنیادی صدا برای مردان ۸۰ الی ۲۰۰ هرتز و برای زنان ۱۵۰ الی ۳۰۰ هرتز می باشد (Ashby & Maidment, 2005). با در نظر گرفتن این

[66] Poon & Ng

مطلب اگر یک بار دیگر به درخت تصمیم گیری بدست آمده رجوع نماییم که صورت ساده شده آن در شکل ۵ قابل ملاحظه می باشد، متوجه می شویم که مشخصه فرکانس بنیادی جهت کلاسه بندی داده ها متضمن شرط بزرگ تر یا کوچک تر بودن از عدد ۰/۱۴ کیلو هرتز می باشد که با توجه به بازه های ذکر شده این مقدار یعنی ۱۴۰ هرتز برای تشخیص صدای مردان و زنان قابل اتکا می باشد.



شکل ۵: صورت ساده شده درخت تصمیم گیری بدست آمده

در نهایت باید اذعان نمود که تشخیص جنسیت گوینده با استفاده از مشخصه های صوتی صدا تنها یکی از کاربردهای مختصر در حوزه علوم آکوستیک می باشد که تکنیک های داده کاوی آن را اجرایی نموده است. مسلماً تجزیه و تحلیل آکوستیک صدا در حوزه های مختلفی از علوم و به ویژه علوم بین رشته ای کاربرد های چشمگیر تری خواهد داشت که امروزه نیز شاهد برخی از آن ها هستیم چنان که به عنوان نمونه می توان به بحث تحلیل احساسات گوینده با استفاده از سیگنال های صوتی و یا شناسایی بیماری های حنجره اشاره نمود که به محققان پیشنهاد می شود با متمرکز کردن پژوهش های خود بر این موضوعات بیش از پیش نقش مهم کاوش در داده های آکوستیک را هویدا سازند.

منابع

1. Ashby M. & Maidment J. (2005). Introducing phonetic science. *Cambridge: Cambridge University Press*.
2. Bharat V., Shelale B., Khandelwal K. & Navsare S. (2016). A review paper on data mining techniques. *International Journal of Engineering Science and Computing,* 6 (5), 6268-6271.
3. Biau G. & Scornet E. (2016). A random forest guided tour. *Test,* 25 (2), 197-227.
4. Buyukyilmaz M. & Cibikdiken A. O. (2016). Voice gender recognition using deep learning. *Advances in Computer Science Research,* 58, 409-411.
5. Fatima & Ikbal Khan J. (2016). Classification of data mining techniques & tools: A survey. *International Journal of Innovative Research and Advanced Studies,* 3 (13), 396-399.
6. Gorade S. M., Deo A. & Purohit P. (2017). A study of some data mining classification techniques. *International Research Journal of Engineering and Technology,* 4 (4), 3112-3125.
7. Hemmerling D., Skalski A. & Gajda J. (2016). Voice data mining for laryngeal pathology assessment. *Computers in Biology and Medicine,* 69, 270-276.
8. Jha A., Dave M. & Madan S. (2016). A review on the study and analysis of big data using data mining techniques. *International Journal of Latest Trends in Engineering and Technology,* 6 (3), 94-102.
9. Maksood F. Z. & Achuthan G. (2016). Analysis of data mining techniques and its applications. *International Journal of Computer Applications,* 140 (3), 6-14.
10. Poon M. S. F. & Ng M. L. (2015). The role of fundamental frequency and formants in voice gender identification. *Speech, Language and Hearing,* 18 (3), 161-165.
11. Provost F. & Fawcett T. (2013). Data science for business: what you need to know about data mining and data-analytic thinking. *O'Reilly Media*.
12. Schapire R. E. (2013). Explaining adaboost. *Empirical Inference,* 37-52.
13. Senthildevi K. A. & Chandra E. (2012). Data mining techniques and applications in speech processing-A Review. *International Journal of Applied Research & Studies,* 1 (2), 1-8.
14. Sharma S. & Mittal H. (2016). Data mining unblocking the intelligence in data. *Journal of Network Communications and Emerging Technologies,* 6 (5), 22-28.