مطالعه ای مختصر در خصوص طبقه بندی توالی

چکیده :

طبقه بندی توالی دارای طیف وسیعی از کاربرد ها نظیر تحلیل ژنوم، بازیابی اطلاعات، انفورماتیک سلامت، امور مالی و تشخیص ناهنجاری ها می با شد. متفاوت از انجام طبقه بندی بر روی بردار ها ی ویژگی، توالی ها معمولا فاقد ویژگی ها ی صریح می باشند.حتی با فنون انتخاب ویژگی پیشرفته، بعدیت ویژگی ها ی بالقوه بسیار بالا است و ماهیت متوالی ویژگی ها ر ا به سختی می توان درک کرد. ا ین موجب می شود تا طبقه بندی توالی و دنباله به یک ویژگی چالش بر انگیز تر از طبقه بندی بردار ها ی ویژگی تبدیل شود. ما اقدام به طبقه بندی توالی از حیث روش ها و حوزه ها ی کاربرد مختلف می کنیم. ما مرور جامعی ر ا در خصوص انواع مختلف مسائل طبقه بندی توالی نظیر طبقه بندی اولیه توالی ها و یادگیری نیمه نظارت شده در زمینه طبقه بندی ها انجام می دهیم.

1- مقدمه

طبقه بندی توالی دار ای طیف وسیعی از کاربرد ها می با شد. در تحقیقات ژنومی، طبقه بندی توالی ها ی پروتین به دسته ها ی موجود بر ای یادگیری وظایف و کارکرد ها ی پروتین استفاده می شود(13). در انفورماتیک سلامت، طبقه بندی سری ها ی زمانی ECG (سری ها ی زمانی ضربان قلب) به ما می گوید که آیا داده ها مربوط به یک فرد سالم هستند یا مربوط به یک بیمار مبتلا به بیماری قلبی است(59). در تشخیص نفوذ و ناهنجاری، توالی فعالیت ها ی دسترسی سیستم بر روی یونیکس بر ای تشخیص رفتار ها ی غیر طبیعی و ناهنجاری پایش می شود(33). در باز یابی اطلاعات، طبقه بندی اسناد به مقوله ها ی موضوعی مختلف، توجهات زیادی ر ا به خود جلب کرده است(51). سایر مثال ها ی جالب شامل طبقه بندی توالی ها ی کوئری بر ای تمایز ربات ها ی ا ینترنتی از کاربر ان انسان(58، 18) و طبقه بندی داده ها ی توالی تر انسفکشن در یک بانک بر ای مبارزه با پول شویی، می با شد(42).

به طور کلی، یک توالی، فهرست منظمی از رویداد هاست. یک رویداد ر ا می توان به صورت یک ارزش نمادین، یک ارزش واقعی عددی، یک بردار با ارزش واقعی یا یک داده نوع پیچیده در نظر گرفت. در ا ین مقاله، داده ها ی توالی یا دنباله ای به زیر انواع زیر در نظر گرفته می شود

- با توجه به القای علایم و نماد ها *fE1; E2; E3; :::; Eng*، یک توالی نمادین ساده ، فهرست منظمی از نماد ها از الفبا می با شد. بر ای مثال، یک توالی DNA متشکل از چهار امینو اسید A-C-G-T و قطعه DNA نظیر ACCCCCGT می با شد که یک توالی نمادین ساده است.
- یک توالی نمادین ساده، فهرستی از بردار ها می با شد. هر بردار یک زیر مجموعه ای از الفبا(34) می با شد. بر ای مثال، بر ای توالی ایتم ها ی خریداری شده توسط یک مشتری در یک سال، در نظر گرفتن هر تر اکنش به صورت یک بردار، یک توالی یا دنباله می تواند به صورت ساعت است (شیر؛ نان) (شیر؛ تخم مرغ) (سیب زمینی؛ پنیر؛ کک) در نظر گرفته شود.
- سری ها ی زمانی ساده دنباله ای از ارزش ها و مقادیر واقعی مرتب شده به ترتیب صعودی می با شد

  h(t1; 0:1)(t2; 0:3) ¢ ¢ ¢ (tn; 0:3)*i*

  یک سری زمانی ساده می با شد که داده ها ی مربوط به مهر زمانی t1 تا t2 ر ا نشان می دهد
- سری ها ی زمانی چند متغیره، دنباله و توالی از بردار ها ی عددی می با شد. بر ای مثال

  h(*t*1*; h*0:1*;* 0:3*;* 05*i*)(*t*2*; h*0:3*;* 0:9*;* 0:8*i*)¢¢¢(*tn; h*0:3*;* 0:9*;* 0:4*i*)*i*

  یک سری زمانی چند متغیره می با شد.
- در بالا، انواع داده ها ی رویداد ها ساده می با شد. در برخی از زمینه ها، نوع داده ها ی رویداد ها ممکن است پیچیده باشند. بر ای مثال، در یک مجموعه داده ها ی پرونده بیمار (http://www.informsdmcontest2009.org/) هر بیمار با یک توالی طولی از ویزیت ها ی بیمارستانی نشان داده می شود. هر باز بینی یا ویزیت یک رویداد است و با اندازه گیری ها ی عددی چندگانه توصیف می شود. یک توالی رویداد پیچیده اشاره به شکل کلی از توالی ها و دنباله ها دارد

  یک توالی یا دنباله ممکن است حامل یک برچسب با شد. بر ای مثال، سری ها ی زمانی داده ها ی ECG بر گرفته از افر اد سالم یا بیمار می باشند. یک توالی DNA مربوط به یک منطقه ژن کد کننده یا غیر کد کننده می با شد. با توجه به ا ین L یک مجموعه برچسب با شد، طبقه بندی توالی ابزاری بر ای یادگیری کلاسیفایر توالی C می با شد، که دنباله S ر ا بر ای یک برچسب کلاس $l \in L,$ در نظر می گیرد و به صورت $C : s \rightarrow l, l \in L$ تعیین می شود.

در طبقه بندی توالی، هر توالی مربوط به تنها یک برچسب کلاسی بوده و توالی کل بر ای دسته بند قل از دسته بندی وجود دارد. هم چنین سناریو ها ی مختلف دیگری بر ای طبقه بندی توالی وجود دارد. بر ای مثال، بر ای توالی یا دنباله علایم یک بیمار در بلند مدت، شر ایط سلامتی بیمار می تواند تغییر کند. هم چنین سناریو ها ی دیگری بر ای دسته بندی توالی وجود دارد. بر ای مثال، بر ای یک دنباله و توالی ای از علایم بیمار در یک دوره بلند مدت، شر ایط سلامتی بیمار می تواند تغییر کند. بر ای توالی استریمینگ، که به صورت توالی نامحدود در نظر گرفته می شود، به جای پیش بینی یک برچسب کلاس، امکان پیش بینی یک توالی از رویداد ها وجود دارد. ا ین مسئله نیز به صورت یک فر ا یند دسته بندی توالی قوی در نظر گرفته شده است. در ا ین مقاله، ما در مورد انواع مختلف طبقه بندی توالی در بخش 3 صحبت می کنیم.

سه چالش مهم دیگر در طبقه بندی توالی وجود دارد. اولا، بیشتر دسته بند ها نظیر درختان تصمیم گیری و شبکه ها ی عصبی، تنها داده ها ی ورودی ر ا به صورت بردار ویژگی ها در نظر می گیرد. دوما، حتی با روش ها ی مختلف انتخاب ویژگی، می تواند یک توالی ر ا به مجموعه ای از ویژگی ها تبدیل می کند. ا ین ویژگی ها بسیار مهم هستند. بعدیت فضای ویژگی بر ای داده ها ی توالی می تواند بسیار بالا با شد و محاسبه می تواند پر هزینه با شد. سوما، علاوه بر نتایج طبقه بندی صحیح در برخی از موارد، امکان دست یابی به دسته بند قابل تفسیر وجود دارد. ایجاد ک دسته بند توالی قابل تفسیر سخت است زیر ا ویژگی ها ی صریح وجود ندارد.

در ا ین مقاله ما مرور مختصری بر روی روش ها ی طبقه بندی و دسته بندی توالی موجود ار ایه می کنیم. چون بیشتر مطالعات فعلی بر دسته بندی توالی سنتی متمرکز هستند، بخش دوم خلاصه ای از روش ها ی اصلی ر ا بر ای ا ین کار ار ایه می کند. در بخش سوم، ما در مورد انواع فعالیت ها ی دسته بندی توالی نظیر استریمینگ دسته بندی توالی و دسته بندی اولیه توالی ها صحبت می کنیم. در بخش چهارم، ما خلاصه ای از دسته بندی توالی ر ا از دیدگاه کاربردی نظیر داده ها ی سری زمانی، داده ها ی متنی و داده ها ی ژنومی ر ا ار ایه می کنیم. در بخش 5 نیز نتیجه گیری ار ایه شده است.

2- روش های طبقه بندی توالی

روش ها ی طبقه بندی توالی ر ا می توان به سه نوع اصلی تقسیم کرد:

- اولین دسته، طبقه بندی مبتنی بر ویژگی است که توالی ر ا به یک بردار ویژگی تبدیل کرده و سپس از روش ها ی طبقه بندی سنتی استفاده می کند. انتخاب ویژگی نقش مهمی در ا ین نوع روش ها ایفا می کند.
- دومین دسته، طبقه بندی مبتنی بر فاصله توالی است. تابع فاصله ای تشابه بین توالی ها ر ا اندازه گیری کرده و کیفیت طبقه بندی ر ا تعیین می کند
- سومین مقوله طبقه بندی مبتنی بر مدل نظیر استفاده از مدل مارکوف پنهان و سایر مدل ها ی اماری بر ای طبقه بندی توالی هاست

در ادامه ا ین بخش ما یک سری روش ها ی معرف ر ا در سه دسته ار ایه می کنیم. برخی از روش ها از چندین مقوله استفاده کرده اند. بر ای مثال، می توان از SVM با استخر اج ویژگی ها یا تعریف شاخص ها ی فاصله ای استفاده کرد. طبقه بندی توالی با استفاده از SVM در بخش 2-3 خلاصه شده است. همه روش ها ی بحث شده در ا ین بخش، طبقه بندی توالی سنتی می باشند.

2-1 طبقه بندی مبتنی بر ویژگی

روش ها ی طبقه بندی سنتی نظیر درختان تصمیم گیری و شبکه ها ی عصبی، بر ای طبقه بندی بردار ها ی ویژگی طر احی می شوند. یک شیوه بر ای حل مسئله طبقه بندی توالی، تبدیل توالی به بردار ویژگی ها از طریق انتخاب ویژگی است

بر ای یک توالی نمادین، ساده ترین ر اه، در نظر گرفتن هر عنصر به صورت یک ویژگی است. بر ای مثال، یک توالی C ACG ر ا می توان به صورت یک بردار ACCG در نظر گرفت. با ا ین حال، ماهیت متوالی توالی ها ر ا نمی توان با ا ین تبدیل توجیه کرد. بر ای حفظ ترتیب عناصر در یک توالی، یک قطعه کوچک از نماد ها ی متوالی K موسوم به K- گرم معمولا به عنوان یک ویژگی انتخاب می شود. با توجه به مجموعه ای از ا ین K – گرم ها، یک توالی ر ا می توان به صورت بردار حضور یا غیاب K- گرم تعیین کرد. گاهی اوقات امکان تطبیق غیر دقیق با K – گرم وجود دارد. با استفاده از ویژگی ها ی K – گرم، توالی ها ر ا می توان با روش طبقه بندی سنتی نظیر SVM تعیین کرد. یک خلاصه ای از روش ها ی انتخاب ویژگی مبتنی بر K – گرم بر ای طبقه بندی توالی ر ا می توان در(16) یافت.

اندازه ویژگی ها ی کاندید که به صورت K – گرم هستند $1 \leq k \leq l \text{ is } 2^l - 1$ می با شد. در صورتی که K یک عدد بزرگ با شد، اندازه ویژگی ها می تواند بزرک تر با شد.چون همه ویژگی ها به طور یکسان بر ای طبقه بندی اهمیت دارند، چانزولو(12) از تست گاما بر ای انتخاب یک زیر مجموعه ای از ویژگی ها با K – گرم استفاده کرده است. یک الکوریتم ژنتیکی بر ای یافتن زیر مجموعه بهینه از ویژگی ها ی محلی استفاده می شود.

بر عکس مجموعه ویژکی ها ی مبتنی بر K- گرم، لیش و همکار ان(30-34) یک روش انتخاب ویزگی مبتنی بر مدل ر ا ار ایه کرده اند. ا ین ویژگی ها به صورت قطعات توالی کوتاهی هستند که معیار ها ی زیر ر ا دارند 1- حدقل در یک دسته دیده می شوند 2- حداقل در یک کلاس یا دسته متمایز هستند 3- اضافی نباشند. معیار(2) به معنی ا ین است که ویژگی بایستی ارتباط و همبستگی معنی داری با یک کلاس داشته با شد. افزونگی در معیار 3 ر ا می توان بر ای طبقه بندی ویژگی و تعمیم ویژگی استفاده کرد. یک الگوریتم ویژگی کاوی کار امد، بر ای کاوش ویژگی ها بر طبق معیار ها پیشنهاد می شود. نتایج ازمایشی در(30) نشان می دهد که مقایسه با روش در نظر گرفتن هر عنصر به عنوان ویژگی، انتخاب ویژگی مبتنی بر الگو موجب بهبود صحت تا 10 تا 15 درصد می شود.

چالش انتخاب ویژگی مبتنی بر الگو در توالی ها ی نمادین ابزاری بر ای جست و جوی ویژگی ها مطابق با معیار هاست. جی و همکار ان یک الکوریتم ر ا بر ای کاوش توالی ها ی با محدودیت فاصله پیشنهاد کرده اند. ا ین الگوریتم که از عملیات بولین و چارچوب رشد پیشوند استفاده می کند، حتی با استانه فرکانس پایین کار امد است سری ها ی زمانی به صورت عددی هستند. روش ها ی انتخاب ویژگی بر ای توالی ها ی نمادین ر ا نمی توان به آسانی به داده ها ی سری ها ی زمانی به کار برد. گسسته سازی می تواند موجب از بین رفتن اطلاعات شود. یی و همکار ان(65) روش انتخاب ویژگی ای ر ا پیشنهاد کردند که قابل کاربرد مستقیم به سری ها ی زمانی عددی است. اشکال مختلف سری ها ی زمانی نشان دهنده یک دسته خاص بوده و به عنوان یک ویژگی بر ای طبقه بندی سری ها مورد استفاده قر ار می گیرد. بر ای یک فر ا یند دو دسته ای، یک واحد به صورت بخشی از سری ها ی زمانی است که بر ای تفکیک داده ها ی اموزشی به دو بخش بر اساس فاصله از واحد استفاده شده و موجب بیشینه سازی بهره وری اطلاعات می شود. استانه فاصله و شکل از داده ها ی اموزشی بر ای بهینه سازی اسفاده می شود. بر ای ایجاد یک دسته بند، فر ا یند انتخاب شکل با ساخت درخت تصمیم گیری تلفیق می شود

اگرچه زیر توالی ها جزو ویژگی ها ی مفید می باشند، با ا ین حال آنها قادر به توصیف ویژگی ها ی محلی با توالی بلند هستند. اگر اول و همکار ان(5) یک روشی ر ا بر ای پوشش دادن ویژگی ها ی جهانی و محلی توالی ها بر ای طبقه بندی توسعه داده اند. اگر اول و همکار ان (5) تجزیه موجک ر ا بر ای توصیف توالی نمادین بر روی وضوح ها ی مختلف اصلاح کرده است. با ضر ایب تجزیه متفاوت، موجک ها بیانکر تغییر ات در دامنه ها و بازه ها ی مختلف چهانی تا محلی است.

با استفاده از تجزیه موجک و دسته بندی مبتنی بر قاعده، روش تجزیه موجک عملکرد بهتری نسبت به دسته بند نزدیک ترین همسایه بر روی مجموعه داده ها ی توالی و مجموعه داده ها ی توالی ژنومی دارد.

به طور خلاصه، روش ها ی موجود از دیدگاه ها ی زیر متفاوت می باشند:

- کدام معیار ها بایستی بر ای انتخاب ویژگی ها یی نظیر منحصر به فرد، فرکانس و طول استفاده شوند؟
- در کدام زمینه انتخاب ویژگی منعکس کننده ماهیت متوالی توالی، محلی و جهانی می باشند؟
- آیا انطباق ها بایستی دقیق یا غیر دقیق با فاصله ها باشند؟
- آیا انتخاب ویژکی بایستی در فر ا یند ایجاد یک دسته بند و یا مرحله پیش پردازش مجزا تلفیق شوند؟

2-2 دسته بندی مبتنی بر فاصله توالی

روش ها ی مبتنی بر فاصله توالی، یک تابع فاصله ای ر ا بر ای اندازه گیری تشابه بین جفت توالی ها ار ایه می کنند. وقتی ا ین توابع فاصله ای انتخاب شدند، می توان از روش ها ی دسته بندی موجود نظیر دسته بند نزدیک ترین همسایه K و SVM با هسته همسوی محلی بر ای دسته بندی توالی استفاده کرد.

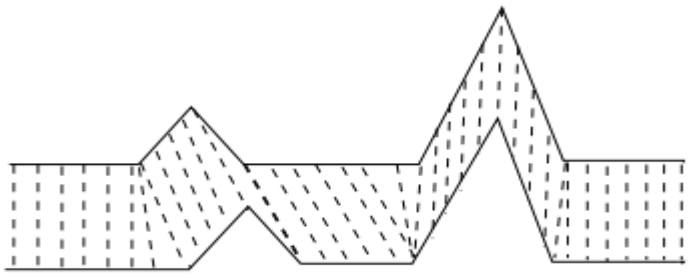
KNN یک روش یادگیری کند بوده و مدل دسته بندی ر ا از قبل محاسبه نمی کند. با توجه به مجموعه داده ها ی توالی T، یک عدد صحیح مثبت K و یک توالی جدید S، دسته بند KNN/، نزدیک ترین همسایه K از s ر ا در T پیدا می کند و سپس برچسب دسته غالب ر ا در KNN به صورت برچسب S قر ار می دهد.

انتخاب شاخص ها ی فاصله ای بر ای عملکرد دسته بند ها ی KNN اهمیت زیادی دارند. در ادامه ا ین بخش، ما بر خلاصه سازی شاخص ها ی فاصله ای مختلف بر ای داده ها ی توالی تاکید داریم. بر ای دسته بندی سری ها ی زمانی ساده، فاصله اقلیدسی استفاده می شود. بر ای دو سری زمانی فاصله افلیدسی به صورت زیر است

$$dist(s, s') = \sqrt{\sum_{i=1}^{L} (s[i] - s'[i])^2}.$$

فاصله اقلیدسی معمولا مستلزم دو سری زمانی با طول یکسان است. کوگ و همکار ان(26) نشان دداده اند که هنگام کاربرد دسته بند 1NN بر روی سری ها ی زمانی، فاصله اقلیدسی از حیث صحت در مقایسه با سایر شاخص ها ی تشابه، رقابتی خواهد بود. فاصله اقلیدسی به شدت به اعوجاج و انحر اف در بعد زمانی حساس است. فاصله تابیدگی زمانی پویا(DTW) بر ای حل ا ین مسئله استفاده شده و نیازمند دو سری زمانی با طول یکسان نیست. ایده DTW همسو سازی سری ها ی زمانی و فاصله ای ار ایه شده است. یک مثال از DTW در شکل 1 نشان داده شده است. زای و همکار ان(61) نشان می دهد که در مجموعه ای از داده ها ی کوچک، شاخص ها ی الاستیک نظیر تابیدگی زمانی پویا صحیح تر از فاصله اقلیدسی است. با ا ین حال نتایج تجربی اخیر نشان می دهد که بر روی مجموعه داده ها ی بزرگ، صحت شاخص ها ی الاستیک با فاصله اقلیدسی هم پوشانی دارد

تابیدگی زمانی پویا با برنامه ریزی پویا محاسبه شده و دار ای پیچیدگی زمانی چند جمله ای است. از ا ین روی، در مجموعه داده ها ی بزرگ پر هزینه است. ر اتامانتا و همکار ان(48) روشی ر ا بر ای تسریع فر ا یند جست و جوی تشابه DTW با استفاده از محاسبات ار ایه کردند. زای و همکار ان(61) از کاهش عددی بر ای تسریع محاسبهDTW استفاده کرده است. ایده اصلی کاهش تعداد مثال ها ی اموزشی مورد استفاده توسط دسته بند 1NN می با شد و در عین حال امکان تنظیم پویای پنجره تابیدگی وجود دارد



شکل 1: تابیدگی زمانی پویا

بر ای توالی ها ی نما دین نظیر توالی ها ی پروتینی و DNA، فواصل مبتنی بر همسو سازی معمولا استفاده می شوند. با توجه به ماتریس تشابه و نیز جریمه فاصله ای، الگوریتم نیدل من یک امتیاز همسو سازی جهانی بهینه ر ا بین دو توالی و دنباله از طریق برنامه نویسی پویا محاسبه می کند. بر عکس الگوریتم ها ی همسو سازی جهانی،

الکوریتم ها ی همسو سازی محلی نظیر اسمیت واترمن و بلاست، تشابه ر ا بین دو توالی با در نظر گرفتن شبیه ترین مناطق اندازه گیری می کنند

3-2 ماشین بردار پشتیبان

SVM یک روش موثر بر ای دسته بندی توالی .[43; 39; 35; 54; 55; 52; 13] می با شد. ایده اصلی استفاده از SVM بر روی داده ها ی توالی، تهیه نقشه توالی بر روی فضای ویژگی و یافتن ابر صفحه با حاشیه حداکثر بر ای تفکیک دو کلاس است. گاهی اوقات، ما نیازی به انتخاب ویژگی صریح نداریم. یک تابع هسته متناظر با یک فضای ویژگی با بعد بالا است. با توجه به دو توالی، X,Y برخی از توابع هسته ای ,K(x; y)، ر ا می توان به صورت تشابه بین دو توالی در نظر گرفت. چالش ها ی استفاده از SVM بر ای طبقه بندی توالی شامل شیوه تعریف فضای ویژگی و توابع هسته و شیوه تسریع محاسبه ماتریس ها ی هسته است

یکی از ر ایج ترین هسته ها بر ای طبقه بندی توالی ، هسته طیف K یا هسته رشته می با شد که توالی ر ا به یک بردار ویژگی تبدیل می کند. لزلی و همکار ان(35) یک هسته طیف K ر ا بر ای طبقه بندی پروتین پیشنهاد کرده اند. با توجه به الفبای امینو اسیدی 20 مولفه ای، ا ین طیف شامل همه دنباله ها ی احتمالی با طول K می با شد که متشکل از عناصر و مولفه ها ی القبا است. بر ای مثال اگر K=3 با شد، طیف K دار ای ARN, AND,DCN می با شد. با توجه به الفبا، A یک دنباله X می با شد که به فضای ویژگی از طریق تابع تر انسفورماسیون تبدیل می شود

$$\Phi_k(x) = (\phi_a(x))_{a \in \mathcal{A}^k}$$

که $\phi_a(x)$ تعداد دفعاتی است که a در x اتفاق می افتد. تابع هسته، محصول نقطه ای بردار ها ی ویژگی است

$$K(x, y) = \Phi_k(x) \cdot \Phi_k(y)$$

با استفاده از الگوریتم درخت پسوند، $K(x, y)$ ر ا می توان در زمان $O(kn)$ محاسبه کرد

لودهی و همکار ان(43) یک هسته رشته ای ر ا بر ای طبقه بندی متنی پیشنهاد کرده است. مشابه با هسته طیف k در(35)، هسته رشته ای از زیر توالی طول k استفاده کرده است. با استفاده از ضریب تجزیه با طول توالی ها در متن، فاصله به صورت جریمه کد گذاری می شود. تابع هسته یک حاصل نقطه ای از بردار ها ی ویژگی بوده و به

طور کار امد از طریق برنامه نویسی پویا محاسبه می شود. لسلی و همکار ان(36) هسته طیف k ر ا بر ای مدیریت عدم انطباق توسعه داده اند. سانبرگ و همکار ان(55) یک هسته با طیف k با عدم انطباق را پیشنهاد کرده اند.

یکی از معایب روش ها ی مبتنی بر هسته ا ین است که تفسیر آن سخت است و کاربر ان به سختی می توان دانش آن ر ا در کنار نتایج طبقه بندی حفظ کنند. سانبرگ یک روشی ر ا بر ای یادگیری svm ها ی قابل تفسیر با استفاده از مجموعه ای از هسته ها ی رشته پیشنهاد کرده اند. ایده اصلی، استفاده از یک ترکیب خطی از اوزان می با شد. هر هسته از یک مجموعه ویژگی ها ی منحصر به فرد استفاده می کند. اوزان بیانگر اهمیت ویژکی هاست. پس از یادگیری svm، کاربر ان می توانند اطلاعات ارزشمندی ر ا در زمینه اهمیت ویژگی ها ار ایه کنند

هسته رشته یا هسته با طیف k ر ا می توان به عنوان یک روش مبتنی بر ویژگی در نظر گرفت. سیگو و همکار ان(49) یک هسته همسوسازی محلی ر ا بر ای دسته بندی توالی پروتین ار ایه کرده است که به عنوان یک روش مبتنی بر فاصله در نظر گرفته می شود. اگرچه فاصله همسو سازی محلی به طور موثر قادر به توصیف تشابه بین توالی هاست، با ا ین حال به طور مستقیم به عنوان تابع هسته ای در نظر گرفته نمی شود زیر ا فاقد ویژگی معین بودن مثبت است. سیگو و همکار ان(49)، فاصله همسویی محلی ر ا اصلاح کرده و یک هسته معتبر ر ا موسوم به هسته همسو سازی محلی ایجاد کردند که از رفتار همسو سازی محلی پیروی می کند. ارتباط نظری بین هسته همسو سازی محلی و فاصله همسو سازی محلی اثبات می شود. با توجه به دو توالی x و y، هسته همسو سازی محلی ر ا می توان با برنامه نویسی پویا محاسبه کرد.

هسته ها ی دیگر مورد استفاده بر ای طبقه بندی توالی شامل هسته ها ی چند جمله ای ، هسته ها ی بر گرفته از مدل احتمال گر ایانه و هسته ها ی انتشار می باشند.

2-4 دسته بندی مبتنی بر مدل

یک مقوله دیگر از روش ها ی دسته بندی توالی، بر مبنای مدل ها ی تولید کننده است که فرض می کند توالی ها در یک دسته از طریق مدل m تولید می شود. با اتوجه به یک دسته ای از توالی ها، M توزیع احتمال اوالی ها ر ا در یک دسته مدل سازی می کند. ا ین مدل بر اساس فرضیات خاص تعریف می شود. در مرحله اموزش، پار امتر ها ی M یاد گرفته می شوند. در مرحله دسته بندی، توالی جدید به یک کلاس با بیشترین احتمال نسبت داده می شود.

ساده ترین مدل، دسته بند توالی بیز ساده می با شد. فرض بر ا ین است که با توجه به یک کلاس یا دسته، ویژگی ها در دنباله ها و توالی ها مستقل از یک دیگر می باشند. احتمالات شرطی ویژگی ها در یک دسته در مرحله اموزشی یاد گرفته می شوند. به دلیل سادگی خود، بیز ساده به طور گسترده توسط دسته بندی تصویر و دسته بندی توالی ژنومی استفاده شده اند

با ا ین حال، فرض مستقل بودن در بیز ساده در عمل نقض می شود. مدل مارکوف و مدل مارکوف پنهان قادر به مدل سازی وابستکی میان عناصر در توالی هاست

یاک شنوف و همکار ان(64) از مدل مارکوف رتبه K بر ای دسته بندی پروتین و داده ها ی توالی متنی استفاده کردند. در فر ا یند اموزش ا ین مدل در شر ایط متفاوت و افتر اقی بر ای افزایش قدرت دسته بندی روش ها ی مبتنی بر مدل اموزش داده می شود. متفاوت از مدل مارکوف، مدل مارکوف پنهان فرض می کند که سیستم مدل سازی شده یک فر ا یند مارکوف با حالت ها ی مشاهده نشده است. سریوستاوا و همکار ان(56) از پروفیل HMM بر ای دسته بندی توالی ها ی زیستی استفاده کرده است. یک پروفیل HMM دار ای سه نوع حالت، افزایش، انطباق و حذف می با شد. مثال ها ی اموزشی بر ای یادگیری احتمال تغییر بین حالت ها و احتمالات حذف استفاده می شود. HMM یاد گرفته شده بیانگر پروفیل مجموعه داده ها ی اموزشی است. یک پروفیل HMM ر ا می توان از توالی ها ی غیر همسو با همسو سازی تدریجی هر مثال با پروفیل موجود یاد گرفت. بر ای هر دسته، یگ پروفیل HMM یادگرفته می شود. در مرحله دسته بندیف یک توالی با HMM در هر دسته با برنامه نویسی پویا همسو سازی می شود یک توالی مجهول به یک دسته با بالاترین امتیاز همسویی دسته بندی می شود.

3- نسخه ها ی مختلف دسته بندی توالی

در ا ین بخش، ما به مرور مسائل مربوط به طبقه بندی و دسته بندی توالی می پردازیم. ا ین مسائل بر ای حل چالش ها در زمان استفاده از طبقه بندی توالی در زمینه ها ی مختلف استفاده می شوند. از ا ین روی از پیشوند ها یی بر ای دست یابی به دسته بندی اولیه، طبقه بندی توالی ها با استفاده از هر دو داده ها ی دار ای برچسب و بدون برچسب و پیش بینی توالی برچسب ها به جای یک برچسب بر ای توالی ها ی استریمینگ استفاده می کنیم

3-1 دسته بندی اولیه

بر ای توالی ها ی نمادین موقت و سری ها ی زمانی، مقادیر توالی به ترتیب صعودی مهر زمانی در نظر گرفته می شود. گاهی اوقات پایش و دسته بندی توالی ها در اسرع وقت مطلوب است. بر ای مثال، در مطالعه کودکان پذیرفته شده در بخش مر اقبت ها ی ویژه، نتایج نشان داد که مودکان دار ای الکوی سری زمانی ضربان قلب غیر عادی بودند. به عنوام مثال دیگرف برنالی و همکار ان(9) نشان داده اند که تنها مشاهده پنج بسته اول TCP ر ا می توان دسته بندی کرد. استفاده از تر افیک انلا ین ر ا می توان بدون انتظار بر ای پایان یافتن چریان TCP شناسایی کرد. تا آن جا که ما می دانیم، دیز و همکار ان(14) اولین بار مفهوم دسته بندی اولیه سری ها ی زمانی ر ا ار ایه کرده اند. آن ها به توصیف سری ها ی زمانی از طریق عبار اتی نظیر افزایش و ماندکاری و معمولا و همیشه پرداختند. به ا ین ترتیب ادا بوست (19) بر ای شبیه سازی دسته بندها ی اصلی استفاده می شوند. ا ین دسته بند قادر به انجام پیش بینی بر روی داده ها ی ناقص با مشاهده پسوند ها ی غیر قابل دسترس توالی ها می با شد

انیبلال و همکار ان(8) از روش استدلال مبتنی بر مورد بر ای طبقه بندی سری ها ی زمانی جهت پایش خر ابی سیستم در سیستم پویا استفاده کردند. دسته بند KNN بر ای دسته بندی سری ها ی زمانی ناقص با فواصل مختلف استفاده می شود نظیر فاصله اقلیدسی و تابیدگی زمانی پویا. مطالعات شبیه سازی با استفاده از استدلال مبتنی بر مورد نشان داده اند که مهم ترین افزایش صحت طبقه بندی بر روی پیشوند ها از طریق سی تا چهل درصد طول کل اتفاق می افتد

اگرچه در(14-8)، اهمیت دسته بندی اولیه بر روی سری ها ی زمانی شناسایی می شود و برخی از نتایج مفید حاصله شده است، ا ین مطالعه دسته بندی اولیه ر ا به صورت دسته بندی پیش وند ها ی توالی ها در نظر می کیرد. زینگ و همکار ان(62) بر چالش ها ی طبقه بندی اولیه تاکید کردند که آن هم مطالعه تعادل بین صحت طبقه بندی و زمان است. روش ها ی پیشنهادی تنها بر انجام پیش بینی بر اساس اطلات جزیی متمرکز است، با ا ین حال به بررسی شیوه انتخاب کوتاه ترین پیشوند بر ای ار ایه پیش بینی مطمئن تاکید دارد. ا ین موجب می شود تا نتایج طبقه بندی اولیه ر ا نتوان به خوبی مورد استفاده قر ار دارد.

زینگ و همکار ان(62) مسئله طبقه بندی اولیه ر ا به صورت طبقه بندی توالی ضمن حفظ صحت مورد انتظار فرموله کردند. یک روش مبتنی بر ویژگی بر ای طبقه بندی اولیه در توالی ها ی نمادین پیشنهاد می شود. ایده اصلی، انتخاب مجموعه ای از ویژگی ها ی منحصر به فرد و سپس ایجاد یک ر ابطه بین دسته بند قاعده و یک

دسته بند درخت تصمیم گیری با استفاده از این ویژگی ها است. در مرحله دسته بندیف یک توالی با همه قواعد و یا شاخه ها به طور هم زمان منطبق می شود تا زمانی که یک پیشوند پیدا شده و توالی طبقه بندی شود. به این ترتیب، یک توالی فور ا زمانی طبقه بندی می شود که صحت ودقت مورد انتظار کاربر حاصل شده باشد. روش ها ی پیشنهادی در 62 برخی موفقیت ها ر ا در مدیریت توالی ها ی نمادین با دست یابی به صحت رقابتی با استفاده از نیمی از طول توالی ها نشان داده اند.

یکی از معایب روش ها در (62) این است که قادر به مدیریت سری ها ی زمانی عددی نیست. چون سری ها ی زمانی عددی بایستی به طور انلاین طبقه بندی شوند، از دست رفت اطلاعات موجب می شود تا نتوان برخی از ویژگی ها ی منحصر به فرد ر ا پوشش داد. زینگ و همکار ان(63) یک دسته بند اولیه ر ا بر ای سری ها ی زمانی عددی با استفاده از یادکیری مبتنی بر فاصله ار ایه کردند. این روش، طول پیش بینی حداقل ر ا بر ای هر سری زمانی در مجموعه داده ها ی اموزشی از طریق خوشه بندی یاد گرفته و از MLP بر ای طبقه بندی استفاده می کند. همان طور که در بخش 2 نشان داده شده است، دسته بند 1NN با فاصله اقلیدسی، یک دسته بند صحیح بر ای دسته بندی سری ها ی زمانی است. یک ویژگی جالب این روش این است که بدون نیاز به صحت مورد انتظار کاربر، دسته بند موجب فعال سازی دسته بندی شده ضمن این که صحت ر ا نیز حفظ می کند.

3-2 دسته بندی توالی نیمه نظارت شده

تعداد داده ها ی بدون شناسه و برچسب بیش از داده ها ی دار ای شناسه است. برخی از داده ها ی بدون شناسه دار ای ویژگی ها ی مشترکی با داده ها ی با شناسه هستند و در عین حال حاوی ویژگی ها ی اضافی ای هستند که توصیف جامع تری ر ا از یک دسته در اختیار می گذارند. از این روی، با استفاده از داده ها ی بدون برچسب و شناسه، یک دسته بند صحیح تر ر ا می توان تولید کرد.

برای طبقه بندی و دسته بندی متن، طیف وسیعی از داده ها ی بدون برچسب و بدون شناسه وجود دارند. نیگام و همکار ان(46) یک روش دسته بندی نیمه نظارت شده ر ا بر ای نشانه گذاری و برچسب گذاری اسناد پیشنهاد کرده اند. در ابتدا،دسته بند ساده بیز بر ای دسته بندی نمونه ها ی بدون برچسب پیشنهاد شد. سپس، فر ایند بیشینه سازی انتظار بر ای تعدیل پار امتر ها ی دسته بند بیز ساده و دسته بندی مجدد داده ها ی بدون برچسب

در یک تکر ار پیشنهاد شد. این فر ایند زمانی به پایان می رسد که نتایج دسته بندی پایدار باشد. یک سند ممکن است متعلق به چندین دسته باشد و یا دار ای برچسب ها ی متعددی باشد.

علاوه بر طبقه بندی متن، زانگ و همکار ان(66) یک طبقه بندی نیمه نظارت شده مبتنی بر HMM ر ا بر ای داده ها ی سری ها ی زمانی پیشنهاد کرده اند. این روش از داده ها ی دار ای برچسب بر ای اموزش پار امتر ها ی اولیه HMM مرتبه اول استفاده کرده و سپس از داده ها ی بدون برچسب بر ای تعدیل مدل در یک فر ایند EM بهره می برد. وی و همکار ان از دسته بند نزدیک ترین همسایه بر ای طبقه بندی سری ها ی زمانی نیمه نظارت شده استفاده کردند. این روش بر ای مدیریت شر ایطی استفاده می شود که در آن ها تنها یک تعداد کمی از داده ها ی دار ای برچسب در یک دسته مثبت موجود می باشد. در مرحله اموزش، در ابتدا، همه داده ها ی بدون برچسب به صورت منفی در نظر گرفته می شود. سپس دسته بند 1NN بر ای دسته بندی داده ها ی بدون برچسب در یک تکر ار استفاده می شود

3-3 دسته بندی توالی با یک توالی و دنباله ای از برچسب ها

همان طور که در بخش 1 بحث شد، بر ای استریمینگ و. ساده سازی طبقه بندی توالیف به جای پیش بینی یک برچسب دسته، پیش بینی توالی برچسب ها مطلوب تر است. کادوس این مسئله ر ا به صورت طبقه بندی توالی قوی در نظر می گیرد ولی یک ر اه حل ر ا بر ای این مسئله ار ایه نمی کند.

یک مسئله مرتبط در پردازش زبان طبیعی، موسوم به توالی ها ی برچسب زنی است. وظیفه این فر ایند برچسب زنی هر عنصر در یک توالی است.بر ای مثال، با توجه به یک توالی، که در آن هر کلمه به صورت یک عنصر در نظر گرفته می شود، برچسب زنی توالی یک کلمه ر ا به یک مقوله نسبت می دهد نظیر هویت، عبارت، فعل یا اسم. یک ر اه حل پیشرفته، در نظر گرفتن برچسب ها ی عناصر در یک توالی مرتبط با یک دیگر است. مسئله برچسب زنی با استفاده از زمینه ها ی تصادفی شرطی حل می شود. این مسئله با سایر روش ها نظیر استفاده از مدل ترکیبی HMM و SVM و شبکه عصبی تکر ای حل شده است

4- کاربرد دسته بندی توالی

دسته بندی توالی دار ای طیف وسیعی از کاربرد ها می باشد. بر ای حوزه ها ی کاربردی مختلف فر ایند دسته بندی دار ای ویژگی ها ی متفاوتی است. در این بخش، ما روش ها ی اصلی به کار برده شده در زمینه ها ی مختلف ر ا خلاصه می کنیم.

1-4 داده های ژنومیک

در سال ها ی اخیر، حجم زیادی ا ز توالی ها ی دی ان ای و پروتین در دیتابیس ها ی عمومی نظیر بانک ژنی، دیتابیس توالی نوکلوتیدی EMBL و دیتابیس پروتین انرتز موجود بوده است. بر ای درک وظایف ژن ها و پروتین ها، دسته بندی توالی توجه زیادی ر ا در تحقیقات ژنومی به خود معطوف کرده است.

روش ها ی مبتنی بر ویژکی به طور گسترده ای بر ای طبقه بندی توالی ژنوم استفاده شده اند. بر ای یک توالی نمادین، ساده ترین ر اه، در نظر گرفتن هر عنصر به صورت یک ویژگی است. بر ای مثال، یک توالی C ACG ر ا می توان به صورت یک بردار ACCG در نظر گرفت. با ا ین حال، ماهیت متوالی توالی ها ر ا نمی توان با ا ین تبدیل توجیه کرد. بر ای حفظ ترتیب عناصر در یک توالی، یک قطعه کوچک از نماد ها ی متوالی K موسوم به K- گرم معمولا به عنوان یک ویژگی انتخاب می شود. با توجه به مجموعه ای از ا ین K – گرم ها، یک توالی ر ا می توان به صورت بردار حضور یا غیاب K- گرم تعیین کرد. گاهی اوقات امکان تطبیق غیر دقیق با K – گرم وجود دارد. با استفاده از ویژگی ها ی K – گرم، توالی ها ر ا می توان با روش طبقه بندی سنتی نظیر SVM تعیین کرد. یک خلاصه ای از روش ها ی انتخاب ویژگی مبتنی بر K – گرم بر ای طبقه بندی توالی ر ا می توان در(16) یافت. اندازه ویژگی ها ی کاندید که به صورت K – گرم هستند$1 \leq k \leq l$ is $2^l - 1$ می با شد. در صورتی که K یک عدد بزرگ با شد، اندازه ویژگی ها می تواند بزرک تر با شد.چون همه ویژگی ها به طور یکسان بر ای طبقه بندی اهمیت دارند، چانزولو(12) از تست گاما بر ای انتخاب یک زیر مجموعه ای از ویژگی ها با K – گرم استفاده کرده است. یک الکوریتم ژنتیکی بر ای یافتن زیر مجموعه بهینه از ویژگی ها ی محلی استفاده می شود. بر عکس مجموعه ویژکی ها ی مبتنی بر K- گرم، لیش و همکار ان(30-34) یک روش انتخاب ویزگی مبتنی بر مدل ر ا ار ایه کرده اند. ا ین ویژگی ها به صورت قطعات توالی کوتاهی هستند که معیار ها ی زیر ر ا دارند 1- حدقل در

یک دسته دیده می شوند 2- حداقل در یک کلاس یا دسته متمایز هستند 3- اضافی نباشند. معیار(2) به معنی ا ین است که ویژگی بایستی ارتباط و همبستگی معنی داری با یک کلاس داشته با شد. افزونگی در معیار 3 ر ا می توان بر ای طبقه بندی ویژگی و تعمیم ویژگی استفاده کرد. یک الگوریتم ویژگی کاوی کار امد، بر ای کاوش ویژگی ها بر طبق معیار ها پیشنهاد می شود. نتایج ازمایشی در(30) نشان می دهد که مقایسه با روش در نظر گرفتن هر عنصر به عنوان ویژگی، انتخاب ویژگی مبتنی بر الگو موجب بهبود صحت تا 10 تا 15 درصد می شود.

دسفنده و همکار ان(13) به مقایسه عملکرد SVM، HMM و KNN بر ای دسته بندی داده ها ی توالی ژنومی پرداختند. ان ها نشان داده اند که SVM عملکرد بهتری از سایرین دارد و انتخاب ویژگی نقش مهمی در تعیین صحت دسته بند ها ی SVM ایفا می کند. شی و همکار ان(52) نیز به این نتیجه رسیده اند که Svm موثر ترین روش بر ای طبقه بندی پروتین است. علاوه بر صحت، چالش ها ی دیگر دسته بندی توالی ژنومی، شامل تسریع طبقه بندی بر ای مدیریت داده ها ی حجیم است.

2-4 داده ها ی سری ها ی زمانی

داده ها ی سری ها ی زمانی نوع مهمی از داده ها ی توالی یا دنباله ای می باشند. در کتابخانه داده ها ی سری ها ی زمانی، داده ها ی سری ها ی زمانی در 22 حوزه از جمله کشاورزی، شیمی، سلامت، مالی، صنعت و غیره جنع اوری شده اند داده ها ی سری ها ی زمانی ucr مجموعه ای از دیتاست ها ر ا به عنوان معیار ارزیابی روش ها ی دسته بندی سری ها ی زمانی ار ایه می کند

بر ای داده ها ی سری ها ی زمانی ساده، بر ای استفاده از روش ها ی مبتنی بر ویژگی، انتخاب ویژگی یک فر ایند چالش بر انگیز می باشد زیر ا ما قادر به شمارش ویژگی ها با داده ها ی عددی نیستیم. از اینر وی روش ها ی فاصله ای بر ای طبقه بندی سری ها ی زمانی استفاده می شوند. نشان داده شده است که در مقایسهب ا طیف وسیعی از دسته ند ها، SVM، HMM و نزدیک ترین همسایه دار ای اهمیت زیادی می باشند. به منظور استفاده از روش ها ی مبتنی بر ویژکی در سری ها ی زمانی ساده، قبل از انتخاب ویژگی، داده ها ی سری ها ی زمانی بایستی به توالی ها ی نمادین از طریق تبدیل نمادین تبدیل شوند. در مقایسه با روش ها ی فاصله ای، روش ها ی مبتنی بر ویژکی موجب تسریع فر ایند دسته بندی شده و یک سری نتایج قابل تفسیر ر ا ار ایه می کند. روش ها ی مبتنی بر مدل بر ای دسته بندی سری ها ی زمانی ساده نظیر HMM استفاده می شو.ند.

طبقه بندی سری ها ی زمانی چند متغیره بر ای تشخیص حرکت و اشاره استفاده شده است. داده ها ی چند ماغیره از طریق مجموعه ای از سنسور ها بر ای اندازه گیری حرکت اشیا در مناطق مختلف تولید می شوند. بر ای دسته بندی سری ها ی زمانی چند متغیره، کادوس و همکار ان یک دسته بند مبتنی بر ویژکی ر ا پیشنهاد کرده اند. برخی از فر ا ئیژگی ها بر ای توصیف روند افزایشی یا کاهشی، سری ها ی زمانی چند متغیره استفاده می شوند. لی و همکار ان روشی ر ا بر ای تبدیل سری ها ی زمانی چند متغیره به یک بردار از طریق تجزیه ارزش عددی و تبدیلات دیکر ار ایه کرده اند.SVM بر ای دسته بندی بردار ها استفاده می شود

3-4 داده ها ی متنی

دسته بندی توالی به طور گسترده ای در بازیابی اطلاعات بر ای دسته بندی متن و اسناد استفاده می شود. ر ایج ترین روش ها بر ای دسته بندی متن شامل بیز ساده، و Svm می باشند. دسته بندی متن دار ای انواع مختلفی می باشد که شامل دسته بندی چند برچسبی، دسته بندی سلسله مر اتبی، و دسته بندی متن نیمه نظارت شده است. شبستانی و همکار ان، یک مطالعه دقیق ر ا بر روی دسته بندی متنی ار ایه کرده اند.

5- نتیجه گیری

در این مقاله ما یک مروری بر طبقه بندی و دسته بندی متن داشتیم. داده ها ی توالی به 5 زیر نوع طبقه بندی شد. روش ها ی دسته بندی توالی در روش ها ی مبتنی بر ویژگی ، روش ها ی مبتنی بر فاصله توالی و روش ها ی مبتنی بر مدل طبقه بندی شد. سپس دسته بندی توالی متعارف نیز بحث شد. در نها یت، روش ها ی مختلف در حوزه ها ی کاربردی مختلف مقایسه شدند

نتایج نشان می دهد که بیشتر کار ها بر دسته بندی توالی ها ی نمادین ساده و سری ها ی زمانی ساده تاکید دارند. اگرچه مطالعات کمی بر روی سری ها ی زمانی چند متغیره وجود دارد مسئله طبقه بندی داده ها ی توالی هنوز حل نشده است. به علاوه بیشتر روش ها به دسته بندی توالی متعارف تخصیص داده شده است. استریمینگ طبقه بندی توالی، اولیه، نیمه نظارت شده و ترکیبی از این مسائل بر روی داده ها ی توالی پیچیده که دار ای کاربرد ها ی زیادی می باشند می توانند چالشی بر ای مطالعات اینده باشند.