عنوان: تشخیص موتیف بین دنباله های زمانی بزرگ مقیاس از نوع ولگشت در حوزه مراقبت های درمانی

نویسند(گان): امیر ادهمیان ناصر پور کیانی

چکیده:

در حال حاضر تجزیه و تحلیل داده های حجیم نقش مهمی در اکثر زمینه های علمی و پژوهشی ایفا می کند. با وجود این که در پژوهش های اخیر، صورت مساله بعضا متداول بوده و قبلا در کار های پیشین معرفی شده، اما پدیده عظیم داده [1] چالش های جدیدی را مطرح می کند که منجر به باز نگری در کارهای قبلی خواهد شد. یک نمونه از کاربرد های آن تحلیل سری های زمانی[2] به صورت بلادرنگ در حوزه مراقبت های درمانی است. مواقعی که داده های ورودی طولانی باشند، رویکرد های سنتی برای ردیابی الگو[3] های پر تکرار، وقت گیر و ناکارآمد هستند. در آن سال ها که این مدل ها پیشنهاد میشد، صورت مساله تک بعدی[4] و تعداد متغیر ها کم بود یا محققین با داده های کم حجم و کوتاه مدت سر و کار داشتند. برخی از ویژگی های این نوع مسایل که استفاده از روش های پردازش عظیم داده را تشویق می کند، عبارت اند از: قابلیت شکسته شدن داده های ورودی به تکه های کوچکتر، پراکندگی[5] داده ها، ... . امروزه مدل های بهبود یافته و ابزار های پردازش عظیم داده باید قادر باشند بر این چالش ها غلبه کنند.

به منظور تشخیص بیماری، حسگر های متعددی در نقاط مختلف بدن مریض تعبیه و داده ها جمع آوری می شود. مجموع آن ها جدولی از داده های چند بُعدی[6] تولید می کند. هر بُعد یک جنبه[7] از مشخصات پزشکی فرد بیمار را بازگو می کند. متخصص با مقایسه ابعاد مختلف از این داده ها تلاش می کند ارتباطی بین آن ها پیدا کند. مثلا بیمار با دیدن عکسی خاص، پلک چشم راستش می زند و هم زمان ضربان قلب اش تندتر می شود. متخصص با کنار هم قرار دادن سری های زمانی حاصل از این چند حسگر و جست و جوی موتیف[8] های مشابه در آن ها، بهتر می تواند برای درمان شخص تصمیم بگیرد. معمولا برای بررسی مجموعه داده های حاصل از این حسگر ها از مدل های طبقه بندی با نظارت[9] استفاده می شود. اگر این مدل ها در مرحله یادگیری، فقط از نتایج خروجی مربوط به یک بیمار به عنوان مجموعه داده های آموزشی[10] استفاده کند، مدل مربوطه بیش از حد بیش برازش [11] خواهد شد. این در حالی است که هر چه قدر مورد های بیشتری را معاینه کنیم و همین طور ابعاد متنوع تری را زیر نظر بگیریم، مدل پیشنهادی ما دقیق تر قادر به تشخیص و طبقه بندی نوع بیماری خواهد بود.

در این پژوهش، با استفاده از پنجره های لغزان[12] زیر دنباله های یک سری زمانی از نوع ولگشت[13] را از هم جدا و هر زیر دنباله را با بکار گیری مدل SAX[14] به موتیف های کوچکتر شکسته و برچسب گذاری می کنیم. سپس با استفاده از ماتریس تصادم[15] و داده کاوی در رشته[16] ای از این حروف تلاش می کنیم تا الگویی های مشابه که در جاهای مختلف سری تکرار شده اند پیدا کنیم. از آن جایی که این زیر دنباله های توزیع شده را می توان به قسمت های کوچکتر تقسیم کرد، از روش نگاشت-کاهش[17] در زیر ساخت Hadoop[18] استفاده می شود. داده های مورد نظر برای آزمایش از سیگنال های ECG[19] واقعی در مجموعه داده UCR[20] دریافت شده است.

---

[1] Big Data
[2] Time Series
[3] Pattern
[4] One Dimensional
[5] Distributed Data
[6] Multi-Dimensional
[7] Feature
[8] Motif
[9] Supervised Classification
[10] Learning
[11] Overfitting
[12] Sliding Window
[13] Random Walk
[14] Symbolic Aggregate approXimation
[15] Collision Matrix
[16] String
[17] Map-Reduced
[18] hadoop.apache.org/
[19] Electrocardiography
[20] http://www.cs.ucr.edu/~eamonn/time_series_data/