موسسه آموزش عالی کاوش

نشریه فنآوری وپژوهش نوین

ارائه راهکاری جهت تشخیص جرائم در داده های بزرگ با استفاده از داده کاوی

عبدالرضا عباسی راد[1]، عباس مدرکی شهد[2]

1- دانشجوی ارشد موسسه آموزش عالی کاوش - محمود آباد abdolreza_abbasirad@yahoo.com

2- استادیار موسسه آموزش عالی کاوش - محمود آباد madraky@yahoo.com

چکیده — امروزه جنایات نوعی مزاحمت اجتماعی هستند و هر جامعه ای از صمیم قلب خواستار رفع آن می باشد. از اینرو بررسی، تحلیل و ریشه یابی جرائم مستلزم بهره گیری از ابزارها و تکنیک های ارائه شده در حوزه فناوری اطلاعات می باشد. داده کاوی بعنوان یک ابزار مهم تحلیلی، می تواند روشی برای افزایش کارایی تحقیقات پلیسی و پیشگیری از جنایات و جرائم باشد. در تحقیق حاضر یک روش پیشنهادی بر مبنای روش خوشه بندی مبتنی بر چگالی و با استفاده از معماری هادوپ برای کمک به روند شناسایی الگوهای جرم و جنایت ارائه دادیم. سپس این روش را به داده های جرم که به صورت فرضی ایجاد کرده ایم، اعمال و نتایج را از لحاظ سرعت اجرا با الگوریتم کامینز مقایسه کردیم. نتیجه مقایسه بیانگر این مطلب است که روش پیشنهادی در اجرا بسیار سریعتر از روش کامینز می باشد. همچنین با توجه به چارچوب مورد استفاده و پشتیبانی از سیستم های توزیع شده در روش پیشنهادی، می توان از آن بعنوان یک سیستم جامع نظارتی درمراکز پلیس بهره برد.

واژه‌های کلیدی — داده کاوی، سیستم اطلاعات مکانی، الگوهای جرم، خوشه بندی

1. مقدمه

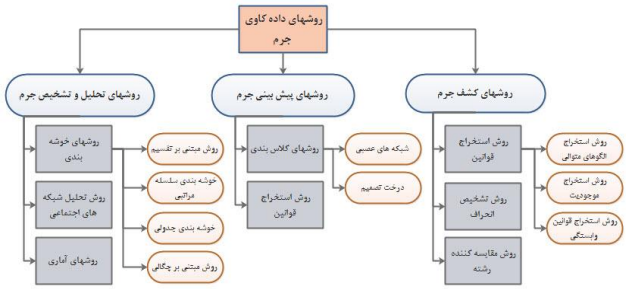
این قالب سالهای زیادی است که جوامع با معضل جرم و چگونگی کنترل آن درگیر بوده اند که تمرکز و خط مشی اصلی آنان در کنترل جرم، تنها افزایش و حرفه ای ترکردن سیستم پلیسی و قضایی بوده است. 1 این روش برای کنترل و مهار جرم در جوامع پر هزینه و غیر منصفانه است. داده کاوی در حوزه جرائم و موضوعات مرتبط به آن و مرور مبانی و چارچوب نظری در زمینه مدیریت جرائم شناخت کلی از ابعاد ان ارائه و درک جنبه های مختلف تحقیق را آسان تر می کند [1]. ایجاد امنیت و آرامش در جامعه تنها با توسل به شیوه های کیفری پس از وقوع جرم محقق نمی شود، بلکه دولت وظیفه دارد با در پیش گرفتن راهکارهایی قبل از وقوع جرم، در از بین بردن شرایط تحقق آن تلاش نماید. لذا پیشگیری از وقوع آن و در صورت وقوع، رسیدگی و کشف آن از مسائل مهم حاکمان و دولتمردان می باشد [2].در برخی کشورها از الگوریتمها و تکنیکهای داده کاوی و ابزارهای آن به منظور تحلیل جرائم استفاده شده است. اما در پلیس ایران تنها از تحلیل های آماری به منظور شناسایی الگوهای جرم و تحلیل جرائم استفاده می شود. همچنین به دلیل عدم وجود سیستم مکانیزه و ثبت اطلاعات مجرمین در سالهای گذشته برای کشف جرائم تنها از کارآگاهان و افسران ماهر و با تجربه استفاده می شد. .. لذا، تحقیق پیشرو سعی در استفاده از برخی الگوریتم های داده کاوی (خوشه بندی) به منظور کشف الگوهای پنهان در پایگاه داده پلیس را دارد [3].

2. پیشینه تحقیق

2.1. روشها و الگوریتمهای مورد استفاده

بطور کلی روشها و الگوریتمهای داده کاوی به سه دسته کلی، اکتشاف جرم، پیشگویی و پیش بینی جرم، آنالیز و تحلیل جرم تقسیم می شوند[1].

شکل 1 بیانگر روشهای مورد استفاده در داده کاوی جرائم می باشد. با توجه به اینکه در این تحقیق فقط بحث تحلیل و تشخیص جرم مطرح است، لذا روشهای موجود دراین حوزه راموردبررسی قرار خواهیم داد.



شکل 1: روشهای داده کاوی جرم

در روشهای تحلیلی که معمولا بر پایه روشهای سنتی آماری، خوشه بندی بدون ناظر و خوشه بندی همراه با ناظر می باشد، داده ها در دسته هایی برای اهدافی خاص طبقه بندی می شوند که این دسته بندی می تواند برای تجزیه و تحلیل به ماموران و کاراگاهان کمک شایانی بکند. این روشها عبارتند از: روشهای آماری ، روش تحلیل شبکه های اجتماعی، روشهای خوشه بندی . روش های خوشه بندی خود به مواردی چون روش های مبتنی بر تقسیم، روش های سلسله مراتبی، روش های خوشه بندی جدولی، روش خوشه بندی مبتنی بر چگالی تقسیم می شوند [4و5].

اگاروال و همکاران در مقاله خود روشی برای خوشه بندی جرائم بر اساس الگوریتم کامینز ارائه کرده اند. ابزار مورداستفاده در تحقیق نرم افزار راپید ماینر بوده است. آنها در این مقاله فقط اقدام به خوشه بندی اقلام داده ای جرائم بر اساس مجموعه داده های متفاوت در سالهای مختلف وصرفا برای جرم " قتل " انجام داده اند. ولی همانطور که میدانیم الگوریتم کامینز بدلیل انجام محاسبات زیاد، زمان اجرای بالایی دارد که در این مقاله اصلا به این موضوع توجهی نشده است. درضمن در همین مقاله نیز به معایب این الگوریتم که برای داده های با نویز بالا و همچنین برای خوشه بندی داده ها در اشکال غیر منظم مناسب نیست، اشاره شده است [6].

2.2. استفاده از داده های بزرگ برای پیش بینی جرائم

پیشرفت های اخیر فن آوری در پردازش و تجزیه و تحلیل داده های بزرگ و پیچیده باعث ارائه فرصت ها و چالش هایی برای پلیس و سازمان های امنیتی شده است. با استفاده از داده های عظیم و متنوع می توان در سه حوزه کلیدی: پیشگیری از جرم، تشخیص جرم و جنایت و امنیت ملی پرداخت. ەمچنین مسائل نظارتی و درک عمومی در مورد حفظ حریم خصوصی، آزادی های مدنی و مزایای اجتماعی آن را نیز پوشش می دهد. [4].

گسترش داده های عظیم و پیچیده ساخت یافته و غیر ساخت یافته به عنوان "کلان داده" نامیده می شود. در حال حاضر پ‌یشرفت توان کامپیوتری، همراه با روش های فنی و روش های جدید برای جذب، پردازش و تجزیه و تحلیل داده های بزرگ و پیچیده (تجزیه و تحلیل ترافیک کلان داده ها)، فرصت های جدیدی را برای استفاده از داده های عظیم برای به دست آوردن درکی از فعالیت های جنایی و استفاده از منابع کارآمد به وجود آورده است [5].

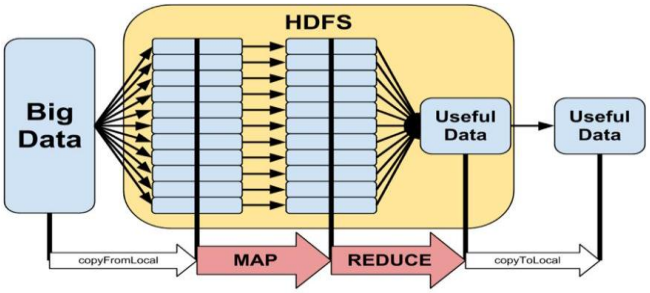
تکنیک های تحلیلی موجود به خوبی در مقیاس های بزرگ کار نمی کند و به طور معمول باعث تولید فاکتورهای مثبت اشتباه می شود که اثر بخشی آنها تضعیف شده است. مشکل زمانی بدتر می شود که سازمان ها رو به معماری ابر و جمع آوری مقدار بیشتری از داده ها می آورند [5].

چارچوب هادوپ (Hadoop) و دیگر ابزارهای کلان داده در حال حاضر منجر به استقرار مقیاس بزرگ و خوشه بندی قابل اطمینان آن شده است که در نتیجه سازمان ها را قادر می سازد تا فرصت های جدیدی برای پردازش و تجزیه و تحلیل داده ها در اختیار داشته باشند [7]. به طور خاص، فن آوری های جدید داده های عظیم، مانند اکوسیستم هادوپ (که شامل پیگ، هایوو ماهوت است) و کاوش جریان، پردازش رویدادهای پیچیده ، و پایگاه داده نواس‌کیوال تجزیه و تحلیل مجموعه داده ناهمگن در مقیاس بزرگ را به طور بی سابقه ای سرعت می بخشد. این فن آوری ها تحولی را در تجزیه و تحلیل امنیتی با تسهیل در ذخیره سازی، نگهداری، و تجزیه و تحلیل امنیت اطلاعات به وجود آورده است [7].

2.3. انتخاب چارچوب برای کار با داده های بزرگ

برای کارکردن با داده های بزرگ ابزارهای مختلفی توسعه داده شده است. شاید بتوان شروع تمام این مباحث را به انتشار مقاله ی شرکت گوگل در مورد مدل برنامه نویسی موازی خود یعنی نگاشت کاهش و هادوپ و جدول بزرگ دانست. بعد از انتشار این مقالات این ابزار به صورت متن باز

در دسترس دیگران نیز گرفته است که تحت لیسانس آپاچی قرار گرفته اند. بعد از این اتفاقات یک سری ابزار دیگر براساس نیازهای دیگر که برمبنای این ابزارها بودند توسط شرکت های مختلف توسعه داده شده اند. از جمله ی این ابزارها می توان به پیگ ، هایو و ... اشاره کرد [8]. چارچوب هادوپ است. معماری کلی این چارجوب در شکل 2 نشان داده شده است.



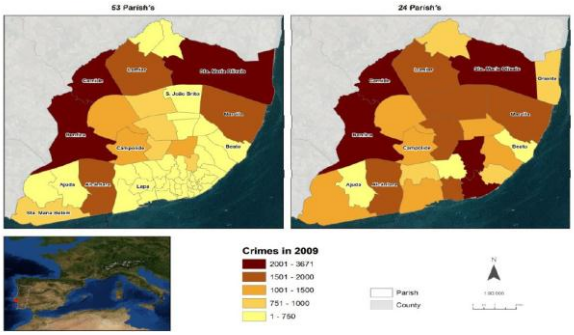
شکل 2: معماری آنالیز داده های بزرگ [8]

2.4. سیستم اطلاعات جغرافیایی جهت تجزیه و تحلیل جرائم

تجزیه و تحلیل جرم به یک مفهوم به اجرای یک نظم و انضباط خاص در جامعه پلیس اشاره دارد. این تجزیه و تحلیل شامل بیش از یک نوع جرم و جنایت می باشد، به همین دلیل است که برخی از نویسندگان به عنوان یک تجزیه و تحلیل ایمنی عمومی به آن رجوع می کنند. با این حال، در طول چند سال گذشته تجزیه و تحلیل جرم و جنایت تبدیل به یک اصطلاح کلی که شامل بسیاری از زیر شاخه های تحقیقاتی است شده است، از جمله: تجزیه و تحلیل اطلاعاتی ، تجزیه و تحلیل تحقیقات جنایی، تجزیه و تحلیل جرم های تاکتیکی، جرم و جنایت راهبردی، تجزیه و تحلیل، تجزیه و تحلیل عملیاتی و تجزیه و تحلیل جرائم اداری [9].

نقشه برداری از جرم و پیرو آن تجزیه و تحلیل مکانی جرم، نقش بسیار مهمی در تعریف شکل های جدیدی از نمایش و تجسم، برای درک بهتر و بیان پاسخ مناسب برای مشکل جنایتکاری دارد. برای درک بهتر عوامل آن، مقامات امنیتی محلی، منطقه ای و ملی به یک ابزار پشتیبانی تصمیم گیری های جدید مانند سیستم های اطلاعات جغرافیایی(GIS) و دیگر فن آوری های اطلاعاتی برای پیدا کردن راه حل های بهتر روی آورده اند. در شهر لیسبون با پیشنهادات جدید بخش اداری، منجر به کاهش از 53 به 24 درجه "freguesias" (حداقل های بخش اداری و اهل محله است) از عدم قطعیت در مشاهدات و محل داده جنایی شده است(شکل 3) [10].شکل 4 نمودار هیستوگرام داده های جرم جمع آوری شده است که نشان می دهد اکثریت جرائم سرقت عمدتا در طول شب اتفاق افتاده است.

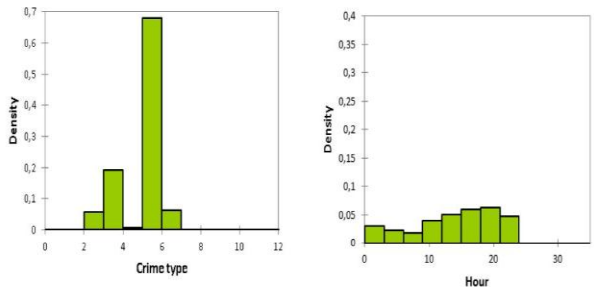
افیانگ و همکاران در مطالعه موردی خود در کشور نیجریه، با بررسی دقیق نقشه های جغرافیایی این کشور و همچنین خوشه بندی انجام شده بر



شکل3: مجموع جنایات در شهر لیسبون [10]

اساس جرائم اتفاق افتاده در مناطق مختلف توانستند محل استقرار مراکز پلیس این کشور را بخوبی تشخیص بدهند. در حقیقت مرکز هر خوشه که بیشترین جرم در آن اتفاق افتاده بر اساس نقشه های جغرافیای مشخص شده و برای احداث مراکز پلیس معرفی شده است [11].

شکل 5 نمایانگر نقشه نقاط جرم در کشور نیجریه است.



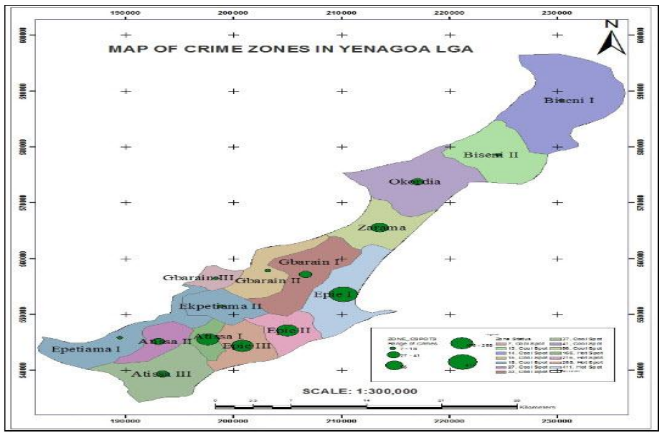
شکل 4 : نمودار هیستوگرام داده های جرم در شهر لیسبون [10]

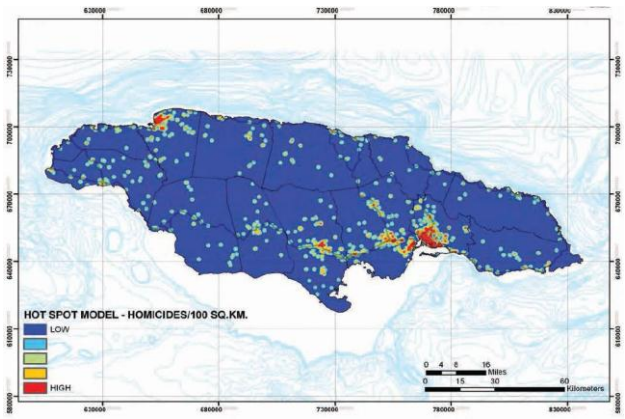
امروزه علاوه بر استفاده از GIS به عنوان نقشه برداری از جنایات و صحنه های جرم و جنایت، پلیس شروع به استفاده از آن در تحقیقات خود نیز کرده است. محققان با استفاده از این ابزارها جهت کمک به یافتن تخصصی روند جرم و الگوها استفاده می کنند، در حالی که هر دو ارزیابی روابط مکانی و زمانی به صورت گسترده ای بر روی داده های جرم، و داده های غیر جرم انجام می شود.

سیاست گذاران از ابزار و استراتژی جغرافیایی برای کمک به تجسم و تجزیه و تحلیل داده های حیاتی مورد نیاز، هزینه، عملکرد و انطباق بهره می برند.

شکل 6 نقشه نقاط جرم در جامائیکا را در سال 2007 نشان می دهد [12].

شکل 5: نقشه نقاط جرم در کشور نیجریه [11]





شکل 6: : نقشه نقاط جرم در جامائیکا در سال 2007 [12]

2.5. بحث و نتیجه گیری

در میان روشهای بکار رفته برای تحلیل جرم دو روش خوشه بندی کامینز و روشهای آماری همچون رگرسیون توسط محققان مورد استفاده قرار گرفته است. لذا به بررسی بیشتر این دو روش خواهیم پرداخت. با توجه به مطالب ارائه شده در همین فصل و با یک جمع بندی کلی می توان مزایا و معایب هر کدام از روشها را بصورت جدول 1 در نظر گرفت. در حقیقت این جدول مزایا و معایب روشهای بکار رفته در تحلیل جرائم را نشان می دهد که در این تحقیق نیز موضوع بحث ما تحلیل جرائم می باشد.

همانطورکه از مطالب جدول مشخص است، با توجه به معایب روشهای سنتی و آماری و عدم پشتیبانی کامل از کلان داده ها و معماری شبکه های توزیع شده، این روشها برای موضوع تحقیق مناسب نیستند. همچنین همین مشکل تا حدی برای روش کامینز هم مطرح است. چراکه این روش نیز دارای زمان اجرای کند بوده و برای اعمال بر روی شبکه های توزیع شده مناسب نیست. اما همانطور که می دانیم ماهیت جرائم با گذشت زمان تغییر می کند مانند جرائم اینترنتی یا جرائم با استفاده از تلفن همراه که قبلا وجود نداشت. بنابراین، به منظور تشخیص الگوهای جدیدتر و ناشناخته در آینده، تکنیک های خوشه بندی بهتر کار می کند. تنها درمقاله اگاروال و همکاران [6] از روش خوشه بندی با استفاده از الگوریتم کامینز برای تحلیل داده های جرم قتل استفاده شده است و دیگر روشها بیشتر برای کشف و پیش بینی جرم بکارگرفته شده است. لذا در این تحقیق که هدف اصلی تحلیل جرم می باشد، روش پیشنهادی با الگوریتم کامینز که توسط محققان برای اینکار استفاده شده بود، مقایسه خواهد شد.

جدول 1: مزایا و معایب روشهای بکار رفته درتحلیل جرائم

نام روش		مزایای روش	معایب روش
روشهای تحلیل جرم	روش خوشه بندی کامینز	نیازبه الگوهای از پیش تعریف شده ندارد	وابستگی به کاربر جهت تعیین مقدار K
		درمقیاس بالای داده (کلان داده) می توان از این روشها استفاده کرد	این روش برای کشف خوشه هایی یا شکل های پیچیده مناسب نیست.
			در پردیس داده های پرت و دوراز مرکز حساس است
			زمان اجرای کند دارد
			عدم پشتیبانی کامل از شبکه های توزیع شده و مسائل مربوط به آنها
	روشهای آماری و سنتی	دسترسی و آشنایی اقراد مختلف در حوزه های مختلف با این روشها	ارحجم بالای داده پشتیبانی نمی کند
		برنامه نهیسی آسانتر	زمان اجرای کند دارد.
			عدم پشتیبانی کامل از شبکه های توزیع شده و مسائل مربوط به آنها

3. روش تحقیق

در این تحقیق به طور خاص از مدل های مبتنی بر خوشه بندی برای کمک به الگوهای جرم و جنایت استفاده خواهیم کرد.

3.1. جمع آوری داده ها و ایجاد مجموعه داده جرم

بیشتر دفاتر کلانتری و ادارات پلیس از سیستم الکترونیکی برای گزارش جنایت به منظور جایگزینی روش کاغذی سنتی استفاده می کنند. ولی به علت محرمانه بودن داده ها و عدم دسترسی مستقیم به آنها، داده های فرضی ایجاد کرده ایم. در ساخت مجموعه داده پروژه سعی بر آن بوده که تمامی

ویژگیهای مربوط به اتفاق یک جرم در نظر گرفته شود که با این حساب 46 ویژگی برای هر جرمی که اتفاق می افتد درنظر گرفته شده است. در واقع مجموعه داده فرضی شامل سه سطر است که در سطر اول نام هر ویژگی وجود دارد. ویژگی هایی شامل ساعت، روز، ماه، سال وقوع جرم، جنسیت مجرم، کدرهگیری، منطقه جرم، طول و عرض جغرافیایی جرم اتفاق افتاده، دستگیر شدن یا نشدن مجرم و ..... در سطر دوم و سوم به ترتیب حداقل و حداکثر مقدار مجاز برای هر ویژگی وجود دارد. نمونه ای از داده ها در جدول 2 نشان داده شده است. بدلیل محدودیت فضای صفحه فقط 10 ویژگی اول در این جدول نشان داده شده است.

جدول 2: قسمتی از یک داده و ویژگیهای آن

Property Name	Hour	Day	Month	Year	CrimeCode	PrimaryType	Location	Arrest	Domestic	Beat
Min Value	0	1	1	1375	1	1	1	0	0	0
Max Value	23	30	12	1396	9999	10	10	1	1	1
Example	3	27	6	1386	30	9	8	0	0	1

یک تابع دیگر نیز به منظور ایجاد ترکیب معناداری از مجموعه داده طراحی شده است. در واقع به این منظور که خوشه بندی معنادار و با هدف باشد باید از هر ویژگی به نسبتی معقول تولید گردد. برای مثال اگر برای هر دو جنسیت مرد و زن به شکل تصادفی و با حضور 50٪ از هر کدام رکورد تولید گردد با داده‌های واقعی تفاوت بسیاری خواهد داشت.

3.2. روش پیشنهادی

در این تحقیق شهر را به صورت یک مربع در نظر می گیریم. هر یک از ابعاد کوچک مربع را بطور جداگانه پردازش می کنیم. به عبارتی شهر رابه تعدادی مربع کوچک تقسیم می کنیم و تک تک این ها را با استفاده از روش نگاشت-کاهش بررسی می کنیم. با توجه به اینکه هر عرض جغرافیایی معادل 111 کیلومتر است، هر ناحیه به طول و عرض 100 متر یک منطقه در نظر گرفته شد. (معادل یک هکتار) با توجه به این که محدوده ی عرض جغرافیایی شهر تهران 51.10 & 51.63 و محدوده ی طول جغرافیایی 35.56 & 35.85 می باشد، در نهایت 350*510 منطقه به دست آمد. اکنون در هر منطقه می توان تعداد جرائم هر ناحیه را مشخص کرد و در نهایت در تابع کاهش با تجمیع نتایج، تعداد جرائم در کل داده ها در هر منطقه مشخص می شود.

سپس چگالی هر یک از مربع ها را حساب می کنیم. چنانچه چگالی آن با هر کدام از همسایه ها مشابه بود، آن دو را به عنوان یک کلاستر در نظر می گیریم. تا جایی که همسایگی هایشان چگالی مشابه نداشته باشند. وقتی یک کلاستر تمام شد به مربع های دیگر می پردازیم. چگالی هر کدام از کلاستر ها به معنای تعداد داده های موجود در هر کلاستر خواهد بود. اکنون با یک عملیات خوشه بندی و با استفاده از یک حد آستانه، نواحی که کمتر از یک حد آستانه در آن ها جرم اتفاق افتاده است، قابل اغماض هستند و مقدار آن ها را صفر می کنیم و در نهایت نواحی همسایه ای که تعداد جرائم آن ها از آستانه ی تعیین شده بیشتر بوده و تنوع جرم در آن ها نیز زیاد باشند، تشکیل یک خوشه می دهند.پرس و جو مورد نظر ما با توجه به هر فیلدی که مورد نظر باشد، می تواند واقع شود و منجر به یک خوشه بندی مفهومی روی محیط گردد. به عنوان مثال می توان تنوع جرائم ویا جرائم بر اساس جنسیت در هر منطقه را تعیین نمود.

به منظور اجرای موازی داده های بسیار بزرگ، و با توجه به محدودیت حافظه ی سیستم و اینکه داده های بزرگ را به صورت یک جا نمی توان در حافظه بارگذاری کرد، الگوریتم نگاشت - کاهش مطرح گردید[13]. در این الگوریتم ابتدا داده ها به صورت تکه تکه خوانده می شوند و هر تکه در اختیار یک تابع نگاشت جهت عملیات قرار می گیرد. پس از انجام عملیات لازم روی هر تکه با استفاده از تابع کاهش نتایج نهایی داده ها جمع آوری می شود و نتیجه ی نهایی استخراج می شود. شکل 8 نشان دهنده فلوچارت روش پیشنهادی می باشد.

شکل 8: فلوچارت روش پیشنهادی

4. ارائه نتایج و ارزیابی عملکرد

الگوریتم پیشنهادی را بر روی دادهای فرضی شهر تهران مورد بررسی قرار خواهیم داد و نتایج را با نتایج بدست آمده از الگوریتم مطالعه مرتبط پیشین (الگوریتم کامینز) مقایسه می کنیم. جهت پیاده سازی الگوریتم و پردازش آن از نرم افزار Matlab 2017a و سیستمی با Intel(R) Core(TM)2 Duo CPU P8600 2.40GHz و 4.00 BG RAM

استفاده کردیم و الگوریتم پیشنهادی و همچنین الگوریتم کامینز را شبیه سازی کرده ایم.

4.1. ارزیابی سرعت اجرای روش پیشنهادی

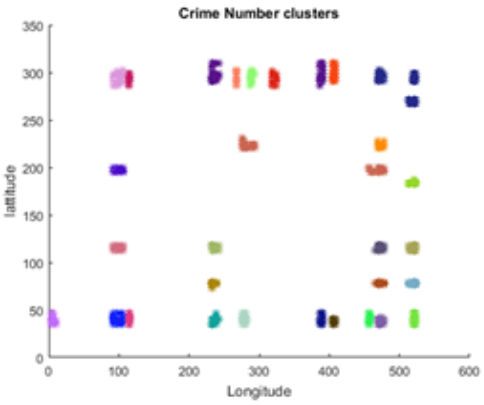
روش پیشنهادی و الگوریتم کامینز را با مقادیر مختلف مجموعه داده مورد بررسی قرار دادیم همانگونه که در جدول زیر مشخص است نتایج زمانی بدست آمده از روش پیشنهادی بر روی دیتاست های متفاوت بسیار مطلوب تر از نتایج بدست آمده از الگوریتم کامینز است که این خود نشان دهنده کارائی و سرعت بالای روش پیشنهادی است.

جدول 3: مقایسه سرعت روش پیشنهادی و K-means

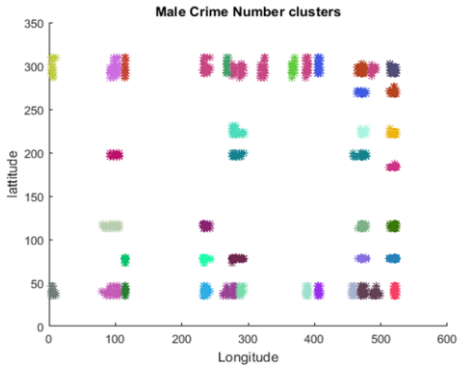
Dataset	*K-means*	*Proposed method*
50000	*27.31*	*15.05*
200000	*134.639*	*27.92*
1000000	*2150.16*	*92.03*
2000000	*Over 7000 S*	*173.38*

4.2. بررسی خوشه بندی های بدست آمده از روش پیشنهادی

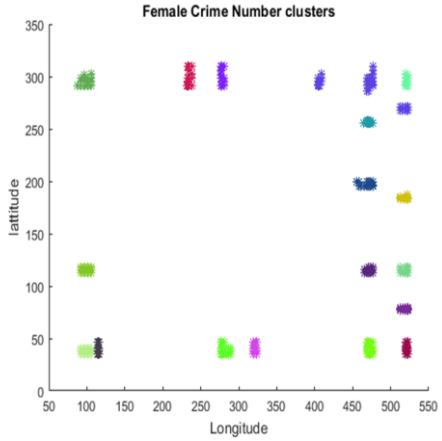
روش پیشنهادی را با مجموعه داده 200 هزار تایی اجرا نموده و خوشه بندی زیر را بدست آوردیم. شکل 9 بیانگر تعداد کل جرائم اتفاق افتاده در طول و عرض جغرافیایی شهر تهران می باشد. در حقیقت محورهای افقی و عمودی این نمودار نشان دهنده طول و عرض جغرافیایی شهر تهران و نقاط رنگی موجود در نمودار نیز نشان دهنده توزیع جغرافیایی جرم در آن منطقه می باشد. شکل 10 نشان دهنده توزیع جغرافیایی جرم در بین آقایان (تعداد مجرم های مرد) و همچنین شکل 11 نیز نشان دهنده توزیع جغرافیایی جرم در بین زنان (تعداد مجرم های زن) در هر منطقه جغرافیایی می باشد. با توجه به این نمودار می توان یک نوع تفکیک جنسیتی برای مناطقی که در آنها جرم اتفاق افتاده را لحاظ کرد. بطور مثال با توجه به نمودار شکل 11 می توان فهمید که معمولا زنان درکدام مناطق جغرافیایی خاص مرتکب جرم می شوند و یا مواردی از این دست. نمودار شکل 12 نشان دهنده انواع جرائم اتفاق افتاده (جرائم هفتگانه) در هر منطقه جغرافیایی می باشد. هر کدام از انواع جرائم با یک رنگ خاص در نمودارمشخص شده اند. جرائم هفتگانه به ترتیب جرائم مربوط به مواد مخدر و الکل (Drug & Alchol)، قتل (Murder)، تجاوز جنسی (Sex Assault)، سرقت (Robbery)، حوادث ترافیکی (Traffic Accidents)، یقه گیری و جیب بری (White Collar Crime) و دیگر جرائم (Others) می باشند.



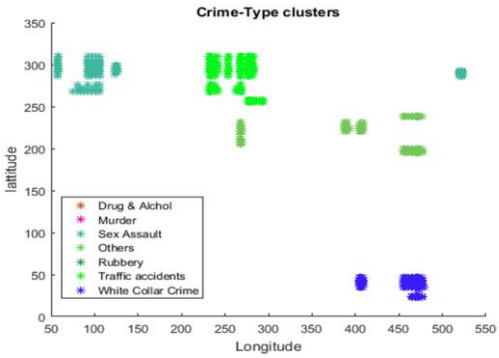
شکل 9: خوشه بندی بر اساس تعداد کل جرائم



شکل 10 : خوشه بندی بر اساس تعداد مجرم های مرد



شکل 11: خوشه بندی بر اساس تعداد مجرمان زن



شکل 12: خوشه بندی بر اساس انواع گوناگونی جرائم

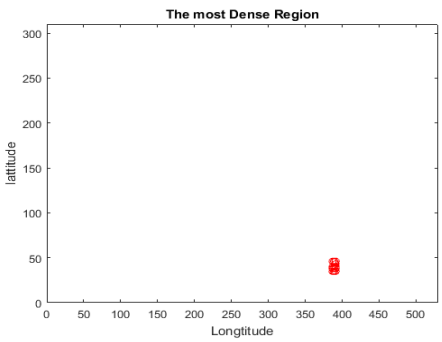
4.3. بررسی چگالی های بدست آمده از روش پیشنهادی

جدول زیر نشان دهنده ی نواحی با بیشترین چگالی جرائم در تهران است. با بررسی این نواحی براحتی مکانهای جرم خیز در شهر تهران مشخص خواهد شد و میتوان اقدامات امنیتی خاص را برای این نواحی در نظر گرفت.در واقع مختصات جغرافیایی نشان داده شده در این جدول نمایانگر مناطق خطرناک و جرم خیز می باشند.

نمودار شکل 13 نشان دهنده طول وعرض جغرافیایی منطقه جرم خیز در شهر تهران با مجموعه داده فرضی است.این نمودار بر اساس چگالی های بدست آمده در جدول 4 رسم می شود.

جدول1: نقاط با بیشترین چگالی جرائم بدست آمده در تهران

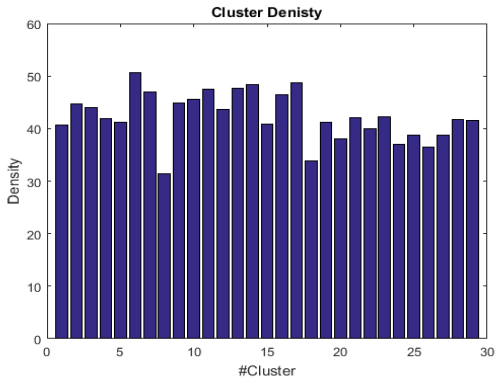
x	Y
35	**387**
35	**389**
35	**391**
38	**387**
38	**389**
38	**391**



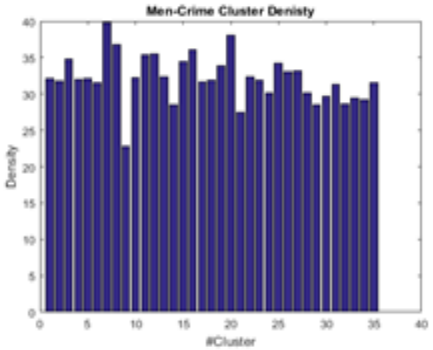
شکل 13 : منطقه جرم خیز

نمودارهای زیر نیز به ترتیب چگالی خوشه ها (میزان جرائم اتفاق افتاده در هر خوشه) در داده های کل، مجرمین مرد و مجرمین زن را نشان می دهد که بیانگر میزان جرم اتفاق افتاده در هر خوشه می باشد.در این نمودارها محور افقی بیانگر تعداد خوشه های خوشه بندی انجام شده و محور عمودی نشان دهنده چگالی یا همان تعداد جرائم اتفاق افتاده در آن خوشه می باشد.

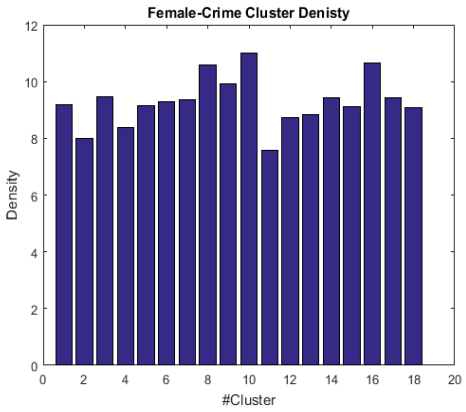
با یک نگاه به نمودارهای شکل 15 و 16 می توان دریافت که داده های فرضی ما تا حد زیادی واقعی می باشند چراکه همانطور که در بحث ساخت مجموعه داده توضیح داده شد، توابع و دستوراتی جهت هرچه واقعی تر شدن داده های فرضی بکار گرفته شد و در آنجا گفته شد که متغیر های مربوط به جنسیت با کمی تغییر در شانس انتخاب روبرو بوده اند و جنبه واقعی بودن متغییر جنسیت این بود که جرائمی که آقایان مرتکب می شوند نسبت به جرائمی که زنان مرتکب می شوند بطور معمول بیشتر است. این مسئله بوضوح در شکل های مربوطه نشان داده شده است. همانطور که مشاهده می شود چگالی جرم آقایان نسبت به چگالی جرم زنان بیشتر است.



شکل 14: توزیع چگالی در خوشه های مبتنی بر داده های کل



شکل 15: توزیع چگالی در خوشه های مبتنی بر مجرم های مرد



شکل 16: توزیع چگالی در خوشه های مبتنی بر مجرم های زن

5. جمع بندی و نتیجه گیری

در این تحقیق از روش خوشه بندی مبتنی بر چگالی برای کمک به تجزیه و تحلیل داده های جرم و جنایت استفاده کردیم. روش مورد نظر مورد تجزیه و تحلیل قرار گرفته و تمامی ابعاد آن در نظر گرفته شد و در نهایت یک روش پیشنهادی بر مبنای روش خوشه بندی مبتنی بر چگالی و با استفاده از معماری هادوپ جهت انجام خوشه بندیهای مورد نظر ارائه دادیم.

نتایج روش پیشنهادی با الگوریتم مورد استفاده پیشین (الگوریتم کامینز) مقاسیه شد و مشخص شد که روش پیشنهادی ما در پایگاه دادهایی با تعداد مختلف سرعت و عملکرد مناسب تری دارد. همچنین با توجه به ماهیت مجموعه داده ایجاد شده و روش پیشنهادی، پرس و جو مورد نظر ما با توجه به هر فیلدی که مورد نظر باشد، می تواند واقع شود و منجر به یک خوشه بندی مفهومی روی محیط گردد. بعنوان مثال در این تحقیق تعداد جرائم اتفاق افتاده در هر منطقه، خوشه بندی بر اساس جنسیت مجرمان مورد بررسی قرار گرفته است.ولی با توجه به ماهیت روش پیشنهادی می توان هر نوع خوشه بندی دیگری را نیز که مد نظر باشد تحقق بخشید.

با استفاده از روش پیشنهادی و با یک محاسبه ی ساده می توان منطقه ی دقیق جرم و محدوده ی طول و عرض جغرافیای آن را شناسایی نمود. همچنین می توان دسته بندی داده ها را بر اساس نوع جرم و یا جنسیت نیز در نظر گرفت و در صورت بروز یک جرم خاص در یک منطقه ی خاص اشخاص مورد نظر را شناسایی نمود. برخی از دستاوردهای مطالعه ما که شامل تجزیه و تحلیل الگوی جرم است می تواند به کارآگاهان در تشخیص هر چه بهتر و سریعتر جرم کمک کند، ولی جایگزین آن ها نیست. همچنین نقشه برداری داده های واقعی در داده کاوی ویژگی است که همیشه کار آسانی نیست و اغلب نیازمند داده کاو ماهر و تحلیلگر جرم با دانش خوب و همکاری نزدیک با یک کارآگاه در مراحل اولیه است.

5.1. پیشنهادها و کارهای آینده

به عنوان کار آینده ، می توان مدلی را برای پیش بینی هات-اسپات های جرم جهت استقرار پلیس در مکان های جرم و جنایت در هر موقعی از زمان طراحی کرد که کمک زیادی به استفاده موثر از منابع پلیس می کند. همچنین با مطالعه به شبکه های اجتماعی توسعه یافته جهت پیدا کردن ارتباط جنایتکاران، مظنونین، باندها و بررسی روابط متقابل آن ها نیز می توان پرداخت. علاوه بر این، توانایی جستجوی مظنون در منطقه، نقض ترافیک پایگاه های اطلاعاتی از کشورهای مختلف و غیره را نیز در نظر داریم. همچنین برای تشخیص هر چه بهتر الگوی جرم مخصوصا اقدامات ضد تروریستی، باید مقادیر و ارزش هایی را به الگوهای تشخیص جرم و جنایت اضافه کنیم.

منابع

[1] Chen, H., et al., *Crime data mining: a general framework and some examples.* computer, 2004. **37**(4): p. 50-56.
[2] Hurwitz, J., et al., *Big data for dummies*. 2013: John Wiley & Sons.
[3] VIOLENCE, A., *Geospatial Technologies and Crime.* Small Arms, 2013. **2**.
[4] Abbi, H. *Big Data, Crime and Security*. July 10, 2014; Available from: http://researchbriefings.parliament.uk/ResearchBriefing/Summary/POST-PN-470.
[5] Daniel E.S. Kawai, D.H.S., *DEVELOPMENT OF CRIMINALS RECORD INFORMATION SYSTEM*, in *NIGERIA COMPUTER SOCIETY (NCS)*. JULY 25-29, 2011.
[6] Agarwal, J., R. Nagpal, and R. Sehgal, *Crime analysis using K-means clustering.* International Journal of Computer Applications, 2013. **83**(4).
[7] Yen, T.-F., et al. *Beehive: Large-scale log analysis for detecting suspicious activity in enterprise networks*. in *Proceedings of the 29th Annual Computer Security Applications Conference*. 2013. ACM.
[8] White, T., *Hadoop: The definitive guide*. 2012: " O'Reilly Media, Inc.".
[9] Corso, A., K. Alsudais, and B. Hilton, *Big Social Data and GIS: Visualize Predictive Crime*. 2016.
[10] Ferreira, J., P. João, and J. Martins, *GIS for Crime Analysis-Geography for Predictive Models.* The Electronic Journal Information Systems Evaluation, 2012. **15**(1).
[11] Effiong, E., et al., *GIS Approach in Analysis of Crime Mapping in Yenagoa Local Government Area of Bayelsa State, Nigeria.* International Journal of Innovative Research and Development, 2016. **5**(10).
[12] Putnik, G.D., *Encyclopedia of networked and virtual organizations*. 2008: IGI Global.
[13] Lu, B. and S. Wei. *One more efficient parallel initialization algorithm of k-means with mapreduce*. in *Proceedings of the 4th International Conference on Computer Engineering and Networks*. 2015. Springer.