داده کاوی فرایند استخراج داده های مفید یا الگو از داده خام می باشد. فرض کنید یک مقدار مشخص از داده داریم، و به دنبال یک حالت مشخص از یک رویداد درون این داده ها می باشیم. می شود انتظار داشت که این رویداد بخصوص رخ دهد، حتی اگر داده ها کاملا تصادفی باشند. تعداد رویداد ها با افزایش اندازه داده ها، افزایش می یابد. اما این روی داد ها فریبنده هستند؛ زیرا داده های تصادفی همیشه تعدادی خصوصیت غیر عادی دارند که به نظر مهم می آیند اما اینطور نیست. یک تئوری آماری به نام تصحیح بنفرونی1، یک روش آماری ارائه می دهد که می تواند اکثر رویداد های فریبنده را تشخیص دهد. بدون ورود به جزئیات آماری، ما یک روش غیر رسمی بر اساس اصل بنفرونی ارائه می کنیم، که کمک می کند به اشتباه رویداد های تصادفی را پردازش نکنیم. با فرض این که داده ها تصادفی هستند، تعداد رخ دادن رویداد مورد نظر را محاسبه می کنیم. اگر عدد به دست آمده به شکل چشمگیری از تعداد واقعی نمومه ها بیشتر باشد، آنگاه باید انتظار داشت هر چیزی که پیدا شده فریبنده باشد. البته این یک مسئله آماری ساختگی است تا یک مدرک برای پیدا کردن چیزی که به دنبال آن هستیم.

داده کاوی در اینترنت اشیاء جهت مدیریت مقدار زیاد داده هایی که توسط دستگاه های اینترنت اشیاء تولید می شوند، کاربرد دارد. داده کاوی شامل کشف و تجزیه و تحلیل دانش از مجموعه گسترده ای از داده ها است. هدف اصلی از استخراج داده، یافتن الگوهای مفید از مجموعه داده های بزرگ دریافت شده از دستگاه های اینترنت اشیاء (IoT) و سنسورها است. کشف دانش، تجزیه و تحلیل الگو ها و برداشت اطلاعات، مفاهیمی هستند که در بحث داده کاوی در اینترنت اشیاء استفاده می شوند. هدف اصلی از داده کاوی، ایجاد یک مدل کارآمد و توصیفی است که برای مجموعه داده مورد نظر مناسب باشد.

II. DATA MINING TECHNIQUES WITH HADOOP AND MAPREDUCE

A. Frequent Itemset Mining (FIM) -FIDOOP

الگوریتم استخراج قانون وابستگی یکی از بهترین الگوریتم های داده کاوی است و برای یافتن مجموعه وقایع پر تکرار و قوانین وابستگی از مجموعه داده های قابل دسترس، مورد استفاده قرار می گیرد.

در قدم اول مجموعه های مکرر شناسایی می شوند و سپس قوانین وابستگی با اعتمادی بزرگتر یا برابر حد اعتماد مشخص شده ایجاد می شوند.

الگوریتم Apriori و الگوریتم FPgrowth دو دسته از الگوریتم استخراج مجموعه وقایع پرتکرار است.

موازی سازی خودکار، توازن بار، توزیع داده ها و تحمل خطا در دسته های بزرگ، نقص اصلی الگوریتم های استخراج موازی موجود برای مجموعه وقایع پرتکرار است.

یک الگوریتم استخراج مجموعه وقایع پرتکرار موازی با نام FiDoop با استفاده از مدل برنامه ریزی MapReduce به عنوان یک راه حل برای این مشکل طراحی شده است.

FiDoop با استفاده از درخت وقایع پرتکرار، به جای درخت طبیعی FP، برای جلوگیری از ایجاد الگوهای نامعلوم و دستیابی به فضای ذخیره سازی فشرده استفاده می شود.

FiDoop سه وظیفه MapReduce را پیاده سازی میکند تا کار داده کاوی خود را تکمیل کند، MapReduce از نقشه گیرها برای جدا سازی مجموعه وقایع مستقل استفاده می کند و کاهش دهنده ها، درخت های کوچک را برای ترکیب عملیات کارهای نقشه گیر mapper ایجاد می کنند.

کلیه وظایف داده کاوی این درخت ها به صورت جداگانه انجام می شود.

در این مقاله، FiDoop با استفاده از خوشه +in-house Hadoop پیاده سازی می شود و از آنجا که طول های مختلف مجموعه وقایع، هزینه های مختلف تجزیه و ساخت را به دنبال دارد، این خوشه به توزیع داده ها و ابعاد آن بسیار حساس است.

تعادل اندازه حجم کار(Workload balance metric) برای اندازه گیری توازن بارگذاری بر روی گره های محاسباتی خوشه ای توسعه داده شده است.

به منظور سرعت بخشیدن به کارکرد استخراج تجزیه و تحلیل داده های با ابعاد بزرگ، یک گسترش از FiDoop به نام FiDoop-HD توسعه داده شده و با مجموعه داده های طیفی آسمانی دنیای واقعی، توسعه یافته و آزمایش شده است.

روش مورد استفاده در آزمایش آنها از مجموعه داده های داده شده M عضوی، یک لیست M- 1 عضوی گرفته و دوباره به M-2 مجموعه تقسیم میکند. این فرآیند تجزیه تا تکمیل مجموعه های پرتکرار نهایی، تکرار میشود.

نتایج آزمایش آن ها نشان می دهد که FiDoop برای مجموعه های بزرگ داده های کم حجم بسیار مناسب است و FiDoop-HD برای پردازش اطلاعات با ابعاد بزرگ بهتر عمل می کند.

FiDoop-HD از تجزیه مجموعه وقایع طولانی برای بهبود عملکرد FiDoop استفاده می کند و برای اندازه گیری تعادل بار FiDoop یک درخت متریک معرفی شده است.

کار آینده این مقاله شامل استفاده از این متریک برای بررسی استراتژی های توازن بار پیشرفته و اجرای طرح متعادل کننده بار با توجه به داده ها در زمینه FiDoop و ادغام مکانیزم قرار دادن داده ها با FiDoop در خوشه های ناهمگن است.برای محاسبه مقدار داده ذخیره شده در هر گره ناهمگن و بهبود عملکرد پردازش داده ها، طرح قرار دادن داده ها استفاده می شود.این استخراج موازی مجموعه های پرتکرار باید بر روی خوشه های ناهمگن با مکانیزم قرار دادن داده ها بررسی شود. [15]

B. FIM on Hadoop

بینش های ارزشمند اقتصادی می تواند توسط مجموعه های پر تکرار فراهم شود و برای اهداف تحقیقی بسیار مفید است.در حالی که استخراج با داده های بزرگ، حریم خصوصی نقش مهمی در داده های غيرمستقيم بدون ساختار دارد.

بنابراین، نحوه محاسبه و حفظ حریم خصوصی مجموعه داده های حاوی اطلاعات حساس یک مسئله بزرگ است.مکانیزم های حفظ حریم خصوصی مختلف برای حفاظت این اطلاعات حساس فراهم شده است.

در این مقاله، نویسنده یک سیستم پیشنهادی که از استخراج مجموعه های پرتکرار (FIM) در MapReduce استفاده می کند و با مجموعه های غیر عددی پردارش انجام میدهد را معرفی کرده است.

الگوریتم رشد FP با استفاده از MapReduce برای دستیابی به همگرایی پردازش داده ها اجرا می شود.سیستم پیشنهادی با چهار مجموعه داده ها و نتایج با سرعت بالا و بهبود استخراج بهره وری آزمایش می شود و می تواند برای استخراج مجموعه داده های فایل های عظیم مورد استفاده قرار گیرد.

در این مقاله، استخراج مجموعه وقایع پرتکرار خصوصی با استفاده از نقشه مدل کاهش برنامه ریزی برای مجموعه داده های بزرگ خصوصی به منظور کاهش زمان مورد نیاز، پیشنهاد می شود. رشد FP بیشتر برای مجموعه داده های عددی استفاده می شود اما سیستم پیشنهادی برای بهبود مکانیزم حفظ حریم خصوصی با استفاده از داده های غیر عددی اجرا می شود. [17]

C. K-Nearest Neighbor

K نزدیک ترین همسایه یک روش معمول طبقه بندی مورد استفاده در داده کاوی است.این روش یک پیوست KNN برای یک مجموعه داده شده از نقاط پرس و جو R و مجموعه ای از نقاط مرجع S را محاسبه می کند و نزدیک ترین همسایه ها را در S برای هر نقطه در R پیدا می کند. هدف اصلی این الگوریتم یافتن پیوست KNN از مجموعه داده های آزمایشی با استفاده از مجموعه داده های آموزشی است. از داده های بدون برچسب ورودی داده شده، یک برچسب کلاس جدید برای نقاط داده جدید کشف می شود و عضویت در کلاس با انجام پرس و جو KNN در مجموعه آموزش محاسبه می شود.

محاسبه KNN همچنین می تواند با روش های دیگر در بسیاری از زمینه ها مانند پزشکی، تجزیه و تحلیل شبکه های اجتماعی، و بیوانفورماتیک و تجزیه و تحلیل سری به کار روند.

دو مشکل اصلی KNN پیوستن حجم داده ها و ابعاد داده است. شاخص ها می توانند برای کاهش تعداد فواصل مورد نیاز برای محاسبه و استفاده از پیش بینی ها برای کاهش ابعاد داده استفاده شود.

با این وجود محدودیت ها در حالی که پردازش KNN روی یک ماشین با داده های بزرگ اجرا می شود افزایش می یابد. KNN را می توان با استفاده از یک برنامه ریزی مقیاس پذیر موازی و توزیع شده با چارچوب Hadoop و یک منبع باز برای پردازش داده ها در مقیاس بزرگ اجرا کرد.

In this paper, the author surveys existing methods for computing KNN on MapReduce both theoretically and experimentally. Three generic steps of KNN computation on map reduce such as data pre-processing, data portioning and computation are used to compare solutions. Each step of experiment is analyzed for load balancing, accuracy and complexity aspects. The author uses a variety of datasets and analyzes the impact of data volume, data dimension and time and space complexity of k and its accuracy. First section of this paper pointed out all possible solutions to follow three main steps to compute KNN over map reduce are preprocessing of data, portioning and actual computation. In each step, the pros and cons of other algorithms also discussed. In the second part of this paper, they performed extensive experiments in the same environment and compared the performance, disk usage and accuracy of all these algorithms. For all algorithms two real datasets with high dimensions are used and it was the first published experiment. For each algorithm, a fine analysis and outlining was performed to find out the importance and difficulty of fine tuning some parameters to obtain the best performance.

این مقاله یک دیدگاه واضح و دقیق از الگوریتم های پردازش KNN را در MapReduce فراهم می کند و همچنین محدودیت های هر الگوریتم را در عمل مشخص میکند. به طور کلی، این مقاله می تواند از دستورالعمل برای انتخاب روش مناسب برای پیوست KNN در MapReduce برای یک مجموعه داده خاص استفاده کند. کارهای بعدی این مقاله، اول پیدا کردن پارامترهای مطلوب برای کاهش تعداد تکرارها و دوم، برای کاهش هزینه های مجدد برای نمایش پرس و جو های پویا است. MapReduce از طریق پیاده سازی Hadoop هنوز یک موضوع باز برای جریان داده ها باقی مانده است. [21]

D. Clustering for Big Data

داده ها به صورت داده های بزرگ با سرعت تکرار در اندازه (حجم) و در فرمت های مختلف (انواع) رشد می کنند. این حجم زیادی از داده ها می تواند از منابع مختلفی باشد یعنی رسانه ها، دستگاه های ارتباطی، اینترنت، کسب و کار و غیره و در عین حال به کارگیری آن، همه با بسیاری از مشکلات و چالش ها مواجه هستند. داده کاوی فرآیندی است که بیشتر در داده های تحلیلی، معمولا داده های کسب و کار یا مرتبط با بازار استفاده می شود - همچنین به عنوان "اطلاعات بزرگ(Big Data)" نامیده می شود. چند تکنیک داده کاوی مانند طبقه بندی، خوشه بندی، تجزیه و تحلیل غلط و استخراج قوانین وابستگی در دسترس هستند. در این مقاله، نویسنده چندین کاربرد و اهمیت روش خوشه بندی را مورد بحث قرار داد. الگوریتم خوشه بندی، ابزای قدرتمند برای بررسی بزرگی حجم داده ها ارائه می دهد. آنها با استفاده از روش های خوشه ای متعددی از جمله روش های سنتی و روش هایی که در این مقاله مورد استفاده قرار گرفتند و با استفاده از مجموعه داده های بزرگ مورد آزمایش قرار گرفتند. نتایج تجربی با مزایا و محدودیت های هر روش در این مقاله ارائه شده است. [14]

E. MapReduce for classification

طبقه بندی یکی از گسترده ترین تکنیک های داده کاوی است و فرآیند توسعه یک مدل است که رکورد ها را در یک مجموعه به کلاس های از پیش تعریف شده اختصاص می دهد. برای توسعه یک مدل طبقه بندی، الگوریتم های مختلف مانند K نزدیکترین همسایه، C4.5، Support Vector Machine و Bayes Naive پیاده سازی شده اند.

هدف اصلی یک مدل طبقه بندی این است که به طور دقیق کلاس هدف برای یک مجموعه داده ورودی مشخص شده، پیش بینی کند. فرایند طبقه بندی با ایجاد مدل بوسیله در نظر گرفتن داده های ورودی به عنوان یک مجموعه داده آموزشی آغاز می شود. MapReduce یک سبک مدل برنامه نویسی است که برای پردازش داده های بزرگ بر روی خوشه های توزیع ماشین های کالا طراحی شده است. این مدل برنامه ریزی MapReduce می تواند برای اجرای یک مدل طبقه بندی استفاده شود. در این مقاله ابتدا الگوریتم های طبقه بندی شده در MapReduce را مورد بررسی قرار گرفته و سپس آنها روش طبقه بندی Bayes Naïve را براساس برنامه ریزی MapReduce معرفی کردند. [22] [23]

F. MapReduce using fuzzy

با توجه به افزایش قابل توجه در اندازه داده ها، انجام تجزیه و تحلیل کارآمد با استفاده از تکنیک های سنتی، مشکل شده است. چندین ویژگی از اطلاعات بزرگ مانند حجم، سرعت، تنوع، متغیر، ارزش و پیچیدگی باعث بسیاری از مسائل باز و چالش های تحقیق میشوند.امروزه ضروری است که تکنیک های داده کاوی کارآمد برای پردازش حجم زیادی داده ها وجود داشته باشد، اما علاوه بر این نیاز به برآورده شدن نیازهای محاسباتی برای پردازش حجم زیادی داده نیز وجود دارد.

The main goal of this research paper is to implement a map reduce framework using fuzzy and crisp techniques, and a comparative study is provided between the results of the proposed systems and the methods that are reviewed in the literature. In this paper, author proposed four proposed systems that are implemented using the map reduce programming model to process on big data. In first system, the fuzzy k-nearest neighbor method as a fuzzy technique and the support vector machine as non-fuzzy technique are the two techniques used in the mapper. In second system, the mode, the fuzzy soft labels and Gaussian fuzzy membership function are the three techniques used in the reducer. Using the fuzzy KNN in the mapper and the mode in the reducer is the first proposed system and Using the SVM in the mapper and the mode in the reducer is the second proposed system and Using the SVM in the mapper and the soft labels in the reducer is the third proposed system, and Using the SVM in the mapper and fuzzy Gaussian membership function in the reducer is the fourth proposed system. The experimental results using different data sets show that the fuzzy proposed methods give a better performance than the crisp proposed method and the method reviewed in the literature.

در این مقاله، یک مطالعه مقایسه ای از تکنیک های طبقه بندی بر روی داده های بزرگ ارائه شده است. دو روش طبقه بندی با الگو MapReduce استفاده می شود؛ k نزدیک ترین همسایه فازی و دستگاه بردار پشتیبانی. چهار الگوریتم پیشنهاد شده دارای دو بخش، نقشه گیر و کاهش دهنده است. در بخش mapper، مجموعه داده ها به توده های بیش از خوشه های محاسباتی تقسیم می شوند و مجموعه ای از سوابق متوسط توسط عملکرد نقشه در قالب یک جفت ((key، value)) ایجاد می شود. نماینده گره های فردی پردازش محاسبات را اجرا می کند و نتایج به عملکرد کاهش می رسد. در قسمت کاهش دهنده، نتایج محاسبات فردی دریافت می شوند و با هم ترکیب می شوند تا نتیجه نهایی را بدست آورند. الگوریتم کاهش دهنده از سه عملکرد استفاده می کند. حالت، برچسب های نرم و فازی گاوس. روش های فازی دقت خوبی درعملکرد نشان می دهند. کار آینده این مقاله، پیاده سازی بیشتری از سایر تکنیک های طبقه بندی با استفاده از الگوریتم های موازی برای بهبود بهره وری استفاده و دقت محاسبات منابع و کاهش زمان اجرای است. [13]

G. A Hybrid Data Mining Algorithm

الگوریتم استخراج داده های ترکیبی

داده کاوی به عنوان یک منطقه فعال تحقیق برای چند دهه گذشته به دست آمده است. در داده کاوی، طبقه بندی، یک تکنیک مهم است که یک نمونه داده را یکی از چندین دسته از پیش تعریف شده تعیین می کند. برای حل مشکلات طبقه بندی، روش های مختلف موفقیت آمیز پیشنهاد شده و آزمایش شده اند. در این مقاله، نویسنده با معرفی ترکیبی از الگوریتم های تکاملی و غیر تکاملی، یک طبقه بندی جدید هیبریدی را معرفی کرد. الگوریتم های برنامه ریزی ژنتیک و تصمیم گیری درخت برای بهبود دقت، قابلیت اطمینان و زمان بندی روش طبقه بندی ادغام می شوند. الگوریتم ترکیبی پیشنهاد شده در این مقاله عملکردی بهتر از استفاده جداگانه درخت تصمیم گیری و برنامه نویسی ژنتیکی را ارائه میدهد. با استفاده از انتخاب ویژگی، دقت بهبود می یابد و با کاهش تعداد ویژگی ها، قابلیت پذیری افزایش می یابد. کار آینده این مقاله مقایسه الگوریتم های دیگر هیبریدی موجود و سپس بهبود الگوریتم ترکیبی پیشنهاد شده برای عملکرد بهتر است. [18] [20]