ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

Analyzing and Mapping the Welfare Schemes for Prisoners after their Prison Period using Machine Learning Algorithms and Big data

Jagadish Kumar, Sophia Kennedy, Rajeswari Nandhakumar, Prarthana Hariharan

Department of Information Technology, Velammal Institute of Technology, Chennai, Tamil Nadu, India

ABSTRACT: This paper mainly focuses to give prisoners a new life after their prison period by providing the opportunities such as further education ,employment and other schemes by analyzing their performance ,quality ,education and their graph of improvement in the prison from the database . The Report generated through this analysis can be used by the prisoners for their development purposes. The idea behind utilizing these opportunities is done by mapping the prisoners personal data, work assessment data, Qualification ,Education level with government schemes such employment and education. The Big data technology and some machine learning algorithms are used . The Analytics part is done by employing certain machine learning algorithms using python. Finally a report is created for each of them using a visualization software called Tableau. By doing so ,a report will be created for each and every person based on the categories and qualifications they fall on. Through this approach ,crimes committed by prisoners again after their release can be prevented to some extent.

KEYWORDS: Big data, Hadoop , Machine Learning algorithms ,Tableau.

I. INTRODUCTION

We are living in a world where crime rates are increasing rapidly day by day. Researchers have proposed several crime prediction and detection techniques to prevent future crimes using demographic data which failed to predict crimes in urban region. There are some cases where the wide ranges of crimes are harder to analyze and predict. Analyzing data undergoes several criteria of crime data from public security data, meteorological data, point of interest, Human mobility, 311 public service complaint data**.** Thus crime prediction and detection methodologies are keeping growing with several advancements in the existing feature.

Our paper is not going to predict or detect crimes whereas it totally about preventing the crimes in the region through providing a different living to prisoners after releasing from prison. Thereby preventing the likelihood of released prisoners committing crime repeatedly. We implement these ideologies through a widely used technology called Big data. Recently Big data is so trending among the nations because of its wide variety of use. Big data was defined to store and process both the structured and unstructured forms of data. The structured form commonly includes the tables, databases, data sets etc.

The unstructured forms of data are the data gathered from social media ,social networks such as Facebook, Twitter, sensor networks , Pinterest, video files ,audio files and so on. These unstructured can also be stored and processed using Big data.

Big data has certain characteristics namely volume , variety , velocity , variability , veracity. Volume is defined as the quantity or amount of storage utilized. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not. Variety is defined as the type and nature of the data. This helps people who analyze it to productively use the resulting insight. Velocity is defined as the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Variability is defined as the Inconsistency of the data set can. Veracity is defined as the quality of data captured and can vary greatly, affecting the accurate analysis.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

II. LITERATURE SURVEY

Crimes are prevalent in our world for centuries. Scientists and researchers have proposed several ideas as papers , journal to predict and detect crimes in our environment. Crimes are classified into many categories from theft, murder to cybercrime. Previously, a crime was detected through the reports of victims. These reports and census from the government were not sufficient enough to predict the future crimes. Each and every idea has its own flaws which lead to insufficient analysis about crime.

There are several methods and algorithms used by various scientists to predict crime but there are some limitations in these methods In the paper published by **HONGJIAN WANG** develops non stationary model for crime rate inference using the urban data. This methodology uses two major criteria's for crime detection such as point of interest and taxi flow of data. The drawback of this technique is it focuses only on certain areas of crime. Thus it cannot predict the major scenes of crime in a city.Certain other researchers such as **Shuai Wang** did crime analysis through visualizing the areas having high frequency of crime through spatial and temporal distribution method. This methodology focuses on finding the nearest neighbour analysis along with the degrees of dispersion and orientation of crime data set. Clusters of crime are visualized through hierarchical spatial cluster tools and space time analysis.

1. Data collection:

According **to OLIVERA KOTEVSKA et al.,[1]** a diverse set of data collected including weather data, crime statistics , demographic data and geospatial data .Weather data is included in previous research has found that temperature influences crime rates [2],demographic data, scalar geospatial data are included to explore the relationship between these city parameters and the resilience network. This method of detecting crime is totally not sufficient for major crime scenes.

2. Data analysis concepts:

According to the paper published by **Sunil Yadav et al[3]** , data mining concept are implemented for creating a regression model for the crimes that had occurred for past 14 years in the country. This methodology used supervised learning, unsupervised learning and semi supervised learning for predicting the accuracy of crime. These patterns of data decrease the performance of the system and efficiency of the result cannot be expected.

3. Machine learning algorithms:

i. **Support Vector Machine**:

In order to predefine the level of crime rate and the given percentage of data**[7]** , Hotspot prediction can use the Support Vector Machine concept. A subset of the crime datasets are selected (percentage or number of data crimes) and classify each of selected data point based on the predefined level of crime rate. The data point which has above than predefined rate are positive or classified as hotspot class and the data point which have below than predefined rate are negative or non-hotspot class. Though this method has been modified, using it is still slow and computationally expensive.

ii. Artificial Neural Network:

One of the proposed method is predicting crime using Artificial Neural Network**[8].** This method is introduced by geographical areas which outperforms traditional policing boundaries for crime prediction. Thus ,ANN can be educated using geographical clusters of crime data. To ease predictive modeling ,ANN predicts accurately more than random work. Inspite of that this method takes long time in training phases.

iii. Decision Tree:

One of the supervised method used to handle the complex classification and produce reasonable classification tree**[9].**The decision tree helps police to discover the crime pattern and predict future trends. This method does not work well on all type of data sets.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

III. TECHNOLOGY FOR DEVELOPMENT

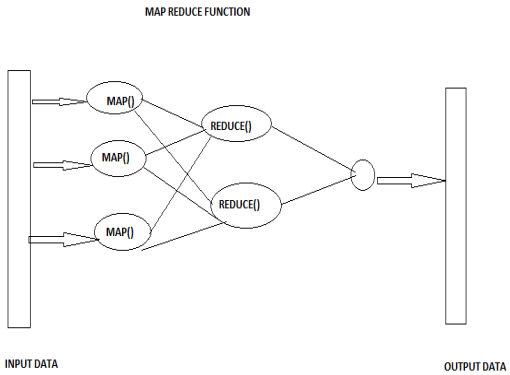
a)Hadoop Framework:.
Apache Hadoop is an open source software framework used for distributed storage and processing of datasets of big data using the MapReduce programming model. It consists of computer clusters built from commodity hardware. The Architecture of Hadoop includes four modules:

- Hadoop common
- Hadoop yarn
- Hadoop Distributed File System
- Hadoop MapReduce

Hadoop common is the java libraries and these libraries will be utilized by the other modules of the hadoop architecture. These libraries are highly needed for starting the hadoop because it contains the java files and java scripts. Hadoop yarn schedules the job in an orderly manner and clustering the resource management. HDFS provides the fastest access for the application data in Hadoop architecture.

b) Hadoop map reduce:
The map task collects all the data and reduces into a key value pair. Conversion from elements to tuples will be done. Reduce task is the later task of the map task. In order to perform the reduce task ,map task has to be done prior. The output of the map task is fed as the input for the reduce task. The map reduce consists of two framework i.e master framework and the slave framework. Master framework is the job tracker which is responsible for the management of resources and checks for the availability of jobs and schedules the job and assigns those jobs to the slave.

An algorithm for Map-Reduce using clusters:



Algorithm works on the principle that making the computer to get the data from where the big data resides .This algorithm mainly has three stages such as map stage , shuffle stage, reduce stage .Initially in the map stage, the data is processed by giving the input data to the mapper function. The mapper function recieves the input line by line. Before providing the data to the mapper function, the data will be retrieved from the Hadoop File System . Then the mapper will divide the given input data into small chunks of data.
In reduce stage ,all processes including the steps of shuffle stage is also performed here. Hence the reduce stage also contains the steps of shuffle stage.

Step 1:During the map-reduce stage of the given algorithm, Hadoop sets the data into maps and reduces the given tasks and assigns it in the appropriate servers in the clusters.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

Step 2:The framework records the information of each activity. The details such as passing of the data , verifying of the task completion and copying the data around the clusters are stored.

Step 3:The computing task takes place in the node. The process that takes place in the local data reduces the network traffic .

Step 4:The cluster collects the data and reduces it and forms the appropriate outcomes and then sends it back to the Hadoop server.

a) Tableau:

Tableau Software helps people to see and understand data. Offering a revolutionary new approach to business intelligence, Tableau allows user to quickly connect, visualize, and share data with a seamless experience from the PC to the iPad. It creates and publishes dashboards and shares them with colleagues, partners, or customers. It does not require any programming skills. Tableau is the most powerful, secure and flexible end-to end analytics platform for data. Tableau is supported on Mac and Windows, with tablet support. Tableau's wide offering of support tools and interactive data analysis makes it easy to understand and learn about trends.

a) Python

Python has gathered a lot of interest recently as a choice of language for data analysis **Tableau Server** is an analytic tool especially for organizations. Dashboards created in Tableau Desktop can be securely shared with Tableau Server. So the whole team can access it from any browser or mobile device and use it. This allows people to communicate with data in a whole new way. It will provide finding answers, solving problems and discover opportunities in smaller time intervals.

Tableau Desktop is an analytic tool for anyone. It is designed to support how people think. Easily combine data by drag & drop to spot trends, identify opportunities, and make data-guided decisions with confidence. Creating interactive presentations allows the audience to explore the data in a self-explaining way. Tableau also allows working offline by extracting data for ad-addhoc analysis of massive data in seconds. It combines adva nces in database and computer graphics technology so it is possible to anale huge datasets on a laptop also.

b) Python

Python has gathered a lot of interest recently as a choice of language for data analysis

- **Scikit Learn** for machine learning. Built on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.
- **Pandas** are used for structured data operations and manipulations. This is extensively used for data munging and preparation. Pandas were added relatively to Python recently and have been instrumental in boosting Python's usage in data scientist community.

c) Machine Learning

In order to make use of machine learning at huge amounts of data or to process stream data, there are three potential levels of impact. First, is to deal with incoming data directly, i.e. data points or dimensions reduction , using sampling-techniques or apply data cleaning. Second, is to alter the processes of data handling, for example putting the processor to the data instead of data to the processor (i.e. a map reduce approach). Finally, is to change implemented algorithms and modify them to fit incoming data. This can be either by directly modifying algorithm procedures or by switching the learning paradigm.

The ability to change or to adapt machine learning algorithms corresponds with the knowledge about what can be changed in the first place. Prior to dive into learning paradigms machine learning techniques can be distinguished from by several features:

- supervised, unsupervised or reinforcement in terms of how learning is achieved
- classification, regression, visualization in terms of what is the result

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

- learning paradigms in terms of how learning techniques are contextually used or combined

It is highly complex to explain the learning paradigms .There are six major paradigms, which may differ in possible aspects they can cope with, in data they are able to process or in the way a model is represented:

1. Deep learning: (deep) neuronal nets are the most common representation. Deep learning produces the representation of data and thus also builds classifications. The major strength of deep learning is to infer features from data.

2. Online learning: Specifically this is able to process streams of data and to update models frequently. There are certain approaches that continually process data and others that make use of micro-batches in order to simulate the continuous processing. Unfortunately, some
of these algorithms are e.g. prone to catastrophic interference.

3. Local learning: Local learning is efficient enough to deal with vast amounts of data. Like in a map reduce way, algorithms or procedures are put in the place where data resides. Data are split into individual segments and partial models are constructed for these segments, which are then taken to build the overall model.

4. Transfer learning: Transfer Learning is a specific approach, which deals with insufficient data for a given domain. In this case, a model is built for one or more other domains where amount of data is sufficient. Additionally, transformations between these domains are defined, so that the generated model is applicable to the domain of interest.

5. Lifelong learning: Lifelong learning can be viewed as a combination of online and transfer learning in that the learning process continually takes new data into account but also provides results for multiple domains. Lifelong learning has the peculiar advantage that the learning process gets more efficient over time.

6. Ensemble learning: This learning technique makes use of several techniques in parallel in order to achieve better results in terms of predictability or accuracy. It can also be used to increase speed of processing. Ensemble learning therefore applies different algorithms on the same data and combines the results in a more or less complex way.

For mapping and analyzing the appropriate welfare scheme for each and every prisoner we use two machine learning algorithms ,they are

i. K-means algorithm:

Machine learning algorithms are mostly used for classification and analysis of data.The main purpose of k-means algorithm is to cluster the data into K-number of cluster

Let us assume inputs as y1,y2,y3…,yn and K value.

- **Step 1** - Select K random points as cluster centers called centroids.
- **Step 2** - Assign each yiyi to the nearest cluster by calculating its distance to each centroid.
- **Step 3**- Take the average of the assigned points do that new cluster center could be found
- **Step 4** - Repeat Step 2 and 3 until none of the cluster assignments change.

Detailed Description of K-means algorithm:

- **Step 1**

We randomly pick **K** cluster centers(centroids). Let's assume these are c1,c2,…,ck c1,c2,…,ck,

C=c1,c2,…, ck

CC is the set of all centroids.

- **Step 2**

In step 2 ,we assign each input value to closest center. It is done by calculating Euclidean(L2) distance between the point and the each centroid. distance.

- **Step 3**

In this step, we find the new centroid by taking the average of all the points assigned to that cluster.

ci=1|Si|∑yi∈Siyi ci=1|Si|∑yi∈Siyi

SiSi is the set of all points assigned to the iith cluster.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

- **Step 4**

In fourth step, we repeat step 2 and 3 until none of the cluster assignments change. That means until our clusters remain stable, repeat the algorithm.

Choosing the Value of K

We often know the value of K. In that case we use the value of K. Else we use the **Elbow Method**. We run the algorithm for different values of **K**(say K = 10 to 1) and plot the K values against SSE(Sum of Squared Errors). And select the value of K for the elbow point **.**

ii. Naive Bayes Algorithm:

Naive bayes is very easy to use and build large data sets. It is helpful in creating sophisticated classification methods. Bayes theorem is a classification technique that uses an assumption of independence among predictors. The assumption is the presence of a particular feature in a class is unrelated to the presence of any other feature. All properties in naive bayes contribute independently to the probability

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Detailed description of Naive bayes:

Step 1: Convert the data set into a frequency table

Step 2: Create Likelihood table by finding the probabilities

Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction. The Naive bayes algorithm are highly used for two major predictions they are

Real time Prediction: Naive Bayes is an eager learning classifier and it is sure fast. Thus, real time predictions could be made.

Multi class Prediction: This algorithm is also well known for multi class prediction feature. Here we can predict the probability of multiple classes of target variable.

iii. Reinforcement learning:

Reinforcement learning is an learning in which it maps the situations to actions. Reinforcement learning use the Markov decision processes framework to define the inter-action between a learning agent and its environment in terms of states, actions, and rewards. This framework simply represent essential features of the artificial intelligence problem. These features include a sense of cause and effect, a sense of uncertainty and non determinism and the existence of explicit goals. The value and value functions are the key features of reinforcement learning methods. The value functions are important for efficient search in the space of policies. The value function distinguishes reinforcement learning methods from evolutionary methods that search directly in policy space guided by scalar evaluations of entire policies.

Trial and error search and delay reward are the two important features of Reinforcement learning. In Reinforcement learning the distinction between the problem and solutions are very important, failing to make this distinction leads to many confusion. Normally the reinforcement learning problem was formalized sing ideas from dynamical system theory, specifically as the optimal control of incompletely known-Markov decision processes. Reinforcement learning is different from Supervised and Unsupervised learning. Reinforcement learning challenges are considered as trade off between the exploration and exploitation. One more feature of reinforcement learning it explicitly consider whole problem of a Goal-directed agent interacting with an uncertain environment.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

1. Markov Decision Process:

The mathematical framework for defining a solution in reinforcement learning scenario is called **Markov Decision Process**. This can be designed as:

- Set of states, S
- Set of actions, A
- Reward function, R
- Policy, π
- Value, V

We have to take an action (A) to transition from our start state to our end state (*S*). In return getting rewards (R) for each action we take. Our actions can lead to a positive reward or negative reward.

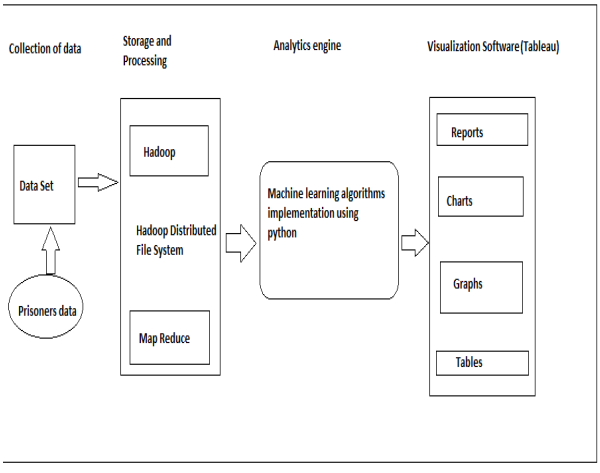
The set of actions we took define our policy (π) and the rewards we get in return defines our value (V). Our task here is to maximize our rewards by choosing the correct policy. So we have to maximize $E(r_t \mid \pi, s_t)$ for all possible values of *S* for a time t.

IV. WORKING PRINCIPLE

In this section ,we introduce a novel architecture for this imagined technology, the big data and its analytics. As a whole ,we have layered the architecture into four major sections, namely 1) Collection of input data 2) Storage and processing using HDFS(Hadoop Distributed File System) 3) Python and machine learning algorithms for analytics 4) visualization using Tableau.

In the first section, the collection of data set is done by gathering the information from the prison department of each state. The data set contains the information about the prisoners personal details along with their activities in the prison. We consider another data set containing the Government schemes that could be provided to the prisoners based on their performance and qualifications.

Second section of the architecture is the Hadoop Distributed File System where the data collected from the prison is stored and processed for further levels of analytics. In HDFS ,Large files are typically distributed in chunks ,and they are stored in data nodes. Applications that run on HDFS have large data set. A typical file in HDFS is gigabytes to terabytes in size. Thus, HDFS is tuned to support large files . It should provide high aggregate data bandwidth and scaled to hundreds of nodes in a single cluster. It should support tens of millions of files in a single instance. Since HDFS provides the fastest access for the application data in Hadoop architecture.



ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

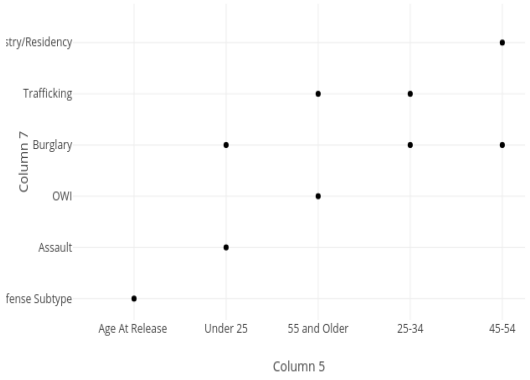
Vol. 7, Special Issue 2, March 2018

The third section is totally about analytics where implementation of machine learning algorithms using python programming language takes place.

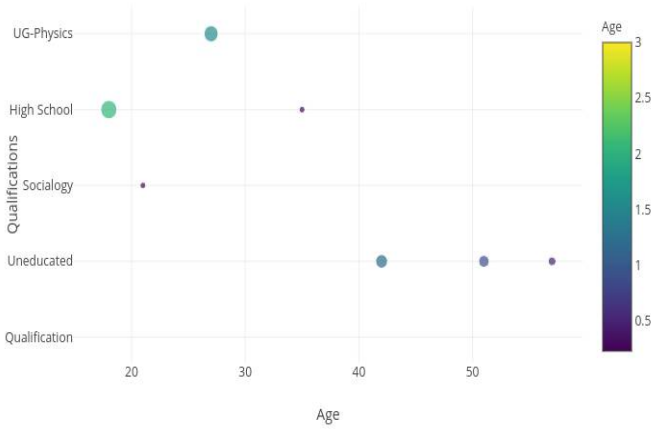
In section four, the visualization software called tableau is utilized for visualizing the report which generated after the deep analysis in the analytics engine. The Tableau is the most powerful , secure and flexible end to end analytics platform for data. It is supported on Mac and Windows with Tablet support.

V. RESULTS

The result of our proposed system is the report that generates for the prisoners according to the categories of schemes that are mapped for their eligibility and qualifications. Since we use naive bayes , k-means and reinforcement learning algorithms ,the process of analytics would be efficient and simple.



The above picture is the graphical representation of prisoners aged under categories and the most likely crime committed by same age group people.



The above graph denotes the qualification of people under certain age. Thus these graph can be mapped together in order to find the best education or employment to the prisoner after leaving the prison.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Special Issue 2, March 2018

VI. CONCLUSION

This paper presents a methodology for providing opportunities to prisoner when they get released from the prison. It is difficult most of the time for prisoners to engage in employment and education after a prisoner tag. Even Job givers are afraid to give them employment. Thus the act of government is important in creating the employment opportunities to the prisoner by providing the report considering their performance , conduct in prison ,education ,ability to do work so that the report will be useful for the prisoner to get their own jobs after their prison period. Thereby there is less likelihood for prisoners to commit crime again since they are provided with some employment or education after their prison period. Our proposed system of analytics uses Hadoop and some simple machine learning algorithms to analyze and map the data and recommends the possible job or stream a prisoner can take. Reports are generated through Tableau software

REFERENCES

[1] OLIVERA KOTEVSKA ,(student member, IEEE),A.GILAD KUSNE, DANIEL V.SAMAROV ,AHMED LBATH,(Member ,IEEE) , and ABDELLA BATTOU "Dynamic Network Model for Smart City Data-Loss Resilience Case Study: City- to-City Network for Crime Analytics" , October 2017.,Vol 5.,pp :29-30
[2] M.Ranson , "Crime ,Weather , and climate change ," J . Environm . Econ.Manage.,vol.67, no.3,pp.274-302,2014.
[3] Sunil Yadav .,Meet Timbadia., Ajit Yadav.,Rohit Vishawakarma , Nikhilesh Yadav ,"Crime Pattern detection ,analysis and prediction.,April 2017
[4] Shuai Wang , Xiaojuan Li , Ying Cai., Jie Tian , " Spatial and temporal distribution and statistic method applied in crime events analysis".,Geoinformatics 2011,19th International conference, June 2011
[5] Hongjian Wang ,Huaxiu Yao ,Daniel Kifer ,Corina Graif ,Zhen Hui Li.,"Non-Stationary Model for Crime Rate Inference using Modern Urban Data". IEEE Transactions on Big Data (Volume:PP,Issue:99) ,December 2017.
[6] Xiangyu Zhao .,Jiliang Tang ,"Exploring Transfer Learning for Crime Prediction",on IEEE International Conference on Data Mining Workshops,2017.
[7] K. Kianmehr and R. Alhajj, "Crime Hot-spots prediction using support vector machine," in IEEE International Conference on Computer Systems and Applications, 2006, pp. 952-959.
[8] M. Chitsazan and G. Rahmani, "Groundwater level simulation using artificial neural network: a case study from Aghili plain, urban area of Gotvand, south west Iran," in Journal of Geopersia, Vol. 3, No. 1, 2013, pp. 35-46.
[9] A. Nasridinov, S.Y. Ihm and Y.H. Park, "A Decision Tree-Based Classification Model for Crime Prediction," Information Technology Convergence, Lecturer Notes in Electrical Engineering, vol. 253, 2015, pp. 531-538.