Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 8, August -2017

Hybrid Technique Based on Clustering for Crime Detection in Data Mining

[1]Chhaya Narwariya, [2]Dr. Shivnath Ghosh

[1]*M.E. (CSE), Maharana Pratap College of Technology Gwalior,*
[2]*Associate Professor, Computer Science & Engineering Maharana Pratap College Of Technology Gwalior*

Abstract— *Crime Detection mainly performed in the Data Mining (DM) to detect the crime efficiently. Crimes are a social annoyance and charge our society extremely in numerous behaviors. Any research that can facilitate in explaining crimes quicker will pay for itself and regarding of this, there are many criminals commit. In the present paper, they implemented a procedure for the design and implementation of crime detection and criminal identification for Indian cities using DM techniques. Clustering is a most vital field of data analysis and data mining application. It is a set of methodologies for creating high superiority clusters and high intra-cluster similarity and low inter-class similarity. In clustering, there is mixture of algorithms to break up the data into groups. K-means is the easiest and most frequently used algorithm for partitioning the data among the clustering algorithms in the field of scientific and industrial software. Fuzzy C-means clustering is utilized to amass the data into groups by characterizing certain degree. We used Fuzzy C-means and ACO in our proposed work to improve the crime detection rate of different cities.*

Keywords— *Crime detection, clustering, classification, Expectation Maximization.*

I. INTRODUCTION

Data mining has been characterized as "the nontrivial extraction of verifiable, beforehand unknown, and conceivably valuable actualities from information. It is “the technology of extracting useful records from massive databases”. Data mining is one of the tasks inside the system of knowledge discovery from the database . Data mining (DM) additionally recognized as Knowledge Discovery in Databases (KDD) has been identified as a rapidly emerging research region. This research area can be distinct as efficiently discovering human knowledge and interesting rules from big databases. DM refers the discovery of exciting know-how, consisting of patterns, institutions, changes, anomalies and substantial systems from large quantities of records saved in databases and different information repositories.[1] There has been an good sized increase in the crime inside the latest past. Crime deterrence has grow to be an upheaval task. The cops of their position to seize criminals are required to stay convincingly ahead within the everlasting race between regulation breakers and law enforcers. One of the key issues of the regulation enforcers is the way to enhance investigative effectiveness of the police.

Crime is depicted as "a demonstration or the commission of a demonstration that is forbidden, or the oversight of an obligation that is directed by means of an open control and that makes the guilty party at risk to discipline with the guide of that law".[2]. Google maps API gives many features for manipulating maps and adding content to the map thru a selection of offerings that permits to create a mapping application. The easily accessible Google Map provides the geometric information for navigation, requiring less memory consumption compared with GIS and Google Street View [3]
Clustering methods changes over dataset to clusters which can be additionally inspected for deciding wrongdoing inclined ranges. These clusters visually represent institution of crimes overlaid on map of police jurisdiction. Clusters keep vicinity of crimes along with other credentials of crime like kind and time. These clusters are labeled on the basis of their individuals. Densely populated clusters end up crime susceptible areas while clusters with fewer individuals are neglected. Preventive measures are implemented in keeping with crime type in crime susceptible areas. K-means is the only and most generally used clustering set of rules in scientific and commercial software. Due to much less computational complexity, it's miles appropriate for clustering large data sets. As such, it has been successfully used in various topics, including market segmentation, computer vision, geo statistics, astronomy and agriculture. It frequently is used as a preprocessing step for different algorithms, as an instance to find a beginning configuration [4].

II. CRIME DETECTION

Data mining is to version crime detection problems. Crime is a social dearly in several ways. Any exploration which could help in illuminating crimes speedier will pay for itself. Crime is considered one of the dangerous elements for any us of a, evaluation of crime is the movement wherein analysis is accomplished on crime sports. Today criminals have most use of all modern-day technology and clinical methods in committing crimes. The law masters should effectually address out difficulties of wrongdoing oversee and upkeep of open request. One test to law implementation and insight organizations is the trouble of breaking down vast volumes of information included in criminal and terrorist exercises.

Thus, making of information base for wrong doings and crooks is required. Information mining holds the guarantee of making it simple, advantageous and reasonable to investigate vast databases for associations and clients.
We have detect the crime on the basis of some criteria-

- Intelligence
- Transference
- Inquisition
- Authoritative.

In crime detection, Crime evaluation is achieved which is defined as analytical approaches which affords relevant data relative to crime patterns and design relationships to help faculty in arranging the sending of assets for the counteractive action and concealment of criminal exercises. It is important to analyze crime due to following motives :

1. Analyze crime to inform law enforcers approximately preferred and particular crime tendencies in well timed manner
2. Analyze crime to take gain of the lots of data present in justice machine and public domain.

Crime rates are rapidly changing and improved analysis finds hidden patterns of crime, if any, without any explicit prior knowledge of these patterns

The essential goals of crime analysis encompass

1. Extraction of crime patterns by means of analysis of available crime and crook facts.
2. Prediction of crime primarily based on spatial distribution of existing data and anticipation of crime charge using distinct data mining techniques
3. Detection of crime [5].

III. CLUSTERING

Clustering is a prime field of data evaluation and data mining application. It is a set of methodologies for produce high superiority clusters and high intra-cluster similarity and low inter -class similarity. The types of data used for analysis of clustering are interval scatted variables binary, nominal, ordinal, ratio variables of mixed types.

A. Hierarchical

A hierarchical method creates a disintegration of the given dataset of the objects hierarchically. Here the tree of clusters called as dendrogram is built. Each cluster node contains child clusters siblings. In hierarchical clustering, we assign.

a) Agglomerative

In agglomerative hierarchical clustering is a bottom up approach each scrutiny starts in its seize cluster, and pairs of cluster are merged as one moves up and about to hierarchy.

b) Divisive

Divisive is a top down approach clustering method all observation starts from one cluster and splits are performed on iterative as one step down the hierarchy unavoidable.

B. Partitioning Based

In the partitioning algorithm split data points in k cluster. The partition in which every point into a cluster and partition is carried out based totally on certain objective feature.

C. Density based

If cluster grows, according to density of neighboring objects and is based on the concept of density reach ability and density connectivity both of which might be relies upon at the input parameter length of the epsilon e neighborhood and minimal phrases of neighborhood distribution of the closest neighbor DBSCAN concentrating on low dimensional spatial data utilized DENCLUE calculation.

D. Grid based

This grid based methods are object space quantizes into a finite number of cells that form a lattice structure. All clustering operations are performed on grid structure, main advantage of this approach is fact process time, which classically independent of the variety of information gadgets and dependent only on the cells in each dimension in Quantize area.

E. High dimensional data clustering

It is a most important tack in cluster analysis because a lot of applications require the analysis of an object. Contain a large number of features or dimension for ex-a text document may contain thousand of keyword as a feature. This over the top dimensional actualities approach of clustering is troublesome because of revile of dimensionality. Many dimensions may not be relevant, the number of dimension gradually more sparse, so that the distance measurement is likely to be low CLIQUE and PROCLUS two subspace clustering method.

F. Constraint based

This constraint based clustering method performs clustering by including user specific or application oriented constraint. A constraint describes user express and provides an effective means for communication with clustering process [6]

IV. K-MEANS CLUSTERING

K- Means is the only and most commonly used partitioning algorithm some of the clustering algorithms in scientific and industrial software program .Acceptance of the K- means is specifically because of its being easy. This algorithm is also appropriate for clustering of the big datasets because it has a lot less computational complexity, though this complexity grows linearly through growing of the data points . Beside simplicity of this approach, it however suffers from a few dangers together with determination of the variety of clusters via user, affectability from outlier records, excessive-dimensional information, and sensitivity in the direction of facilities for preliminary clusters and thus possibility of being trapped into local minimum may additionally reduce performance of the K- means algorithm. K-means clustering is one of the methods of cluster assessment which goals to parcel n perceptions into k clusters wherein each perception belongs to the cluster with the nearest mean.

Process

i. Initially, the quantity of clusters have to be recognised permit or not it's k.
ii. The initial step is the select a set of K times as centres of the clusters.
iii. Next, the algorithm considers each example and assigns it to the cluster that is closest.
iv. The cluster centroids are recalculated both after whole cycle of re-task or every example assignment.
v. This process is iterated.

To address noisy data and exceptions and now not appropriate to discover clusters with non- convex shapes. [7].

V. K-NEAREST NEIGHBOUR (K-NN)

The k-nearest neighbour (k-NN) rule is one of the most popular classification techniques utilized in pattern reputation . An unknown sample is classed as most people magnificence of its okay nearest neighbours within the training set. The success of the k-NN rule depends on which patterns are considered to be the least distant from the unknown pattern. Therefore, the performance of k-NN is determined by the distance definition adopted to measure the closeness. This explains the increasing interest in finding variants of the Euclidean distance to improve the recognition rate of k-NN . In this framework, we reformulate the kNN rule so as to exploit the similarity based on the semantic affinity produced by the neural network described in Section 2. More precisely, let { } N = 1,..., xxT be the training pool. For each element q x of the test set we calculate the similarity degree qj s , between q x and each element j x , j=1..N, in T. We order these similarities according to decreasing values and calculate the probability that element q x belongs to class c as follows:
K manner algorithm complexity is O(tkn), wherein n is times, c is clusters, and t is iterations and comparatively efficient . It often terminates at a neighborhood most beneficial. Its disadvantage is relevant best when imply is defined and need to specify c, the number of clusters, earlier. It unable.

$$P(c) = \frac{\Sigma_{x_{f \in k_q^c}} s_{j,q}}{\Sigma_{x_{j \in k_q}} sj,q}$$

where Kq is the set of elements with the highest similarity degree with q x , and c Kq is the subset of Kq is the subset of Kq containing components having a place with a similar class c [8]

VI. LITERATURE SURVEY

Neha Agarwal(2016)et al presents about Delay Tolerant Network overwhelms the many issues that exist in traditional network. MANET works only in the ones regions where the communique link is to be had at some point of the transferring of packets from supply to destination, while DTN is used in those areas wherein link isn't always properly go out together with in disaster, in rural regions. We proposed an efficacious prophet routing set of rules a good way to supply the trump efficiency and further the buffer length. In this proposed we are able to create a kiosk with a purpose to assist the sharing of crime data in rural regions. This sort of stand is executed where the activity recurrence is high when vehicles will go from this roadway, at that point the high frequency sensor will naturally detect the network and come into their range for transferring the info from rural areas to urban areas. we implemented this concept on the Helensiki road map.[9]

K.Veena(2016)et al presents about The world is facing a lot of issues related to cyber crime. The technology has developed and it should be used for a constructive purpose. But the technology is being used for detrimental purpose and hence we are able to see a whole lot of cyber crime sufferers. As per the statistics gathered inside the beyond years the

cyber crime inside the globe has accelerated to a exquisite extent and the humans committing the crime are using era smartly. We need a safe environment and we want to give a safe environment to the next generation.. It is now most common for any person to try to hack the information we have been using in the internet. Much of training and intelligence is not required to commit the crime. A little bit of technical expertise delivered with a few smartness and a weak man or woman may be easily targeted [10].

Vukosi Marivate(2016)et al presents about The increasing prevalence of Social Media platform use has brought with it an explosion of new user generated public data. This data is centered around many, diverse topics. One theme of interest is tap into the general public protection and crime associated user generated statistics to higher understand patterns inside the prevalence of crime incidents. One challenge in such data is that most of the data needs human annotation to make it usable by machines to analyse. This paper explores how different features, extracted from social media data, impact the performance of different classifiers. The classifiers are built to classify social media data as having to do with a mentioned crime or no longer. The challenge of few labeled data is discussed as well as different approaches to extracting features from the text data as well as the graph created by users associating with each other is investigated [11]

Kalyani C S (2016) et al presents about K-means clustering, averaging followed by morphological operations- convex hull, opening, Last became used to phase rectum from CT images of male pelvic area. The visualization changed into done using most depth projection approach. The proposed method is nearly correct, easy and efficient. The bounding box representation of ROI will decrease the processing time for volume rendering and hence proves to be efficient over the conventional method for utilizing the whole image [12]

Abba Suganda Girsang(2016) et al presents About the algorithm to maintain the best fitness value, this paper presents a model robust algorithm for solving the problem of clustering, namely Robust Adaptive Genetic KMeans, Algorithm, called RAGKA has been shown in this paper. At first solving the problems of clustering using K-Means, but K-Means are often trapped local optimum, then K-Means combined with GA. When using the K-Means with GA still has shortcomings because the optimal results can still be replaced because the probability of crossover and mutation fixed. Therefore, we are trying to use the GA that maintain best fitness value (AGA), and the results are starting to improve, but for some of data is still not optimal, and therefore we try to maintain the best fitness value of the results of the iteration with greedy selection, and the result was increasing rapidly. The results can be seen in Table III, RAGKA to have an accuracy of better than GA with K-Means and AGA with K-Means, because RAGKA maintain the best fitness value. For further research every best fitness value should be maintained, because RAGKA have optimal value by maintaining the best fitness value.[13]

Taranjot Kaur Bajwa(2016)et al presents about Gait recognition is advanced form of biometric technique that is used to identify a person from far away without the knowledge of that person. Gait recognition has overcome over other forms of recognition techniques like face, finger or iris due its improved technology. This is the reason it is used today in military establishment, bank and airports for security purpose. Gait is the manner or style of walking of human being. It is used to recognize or identify a single person or multiple people at the same time. In this paper we are going to analyze Gait to identify person by using SVM with K-NN and NN. In this paper first video of a person is captured and divided into frames. The frames are pre-processed, and background unwanted part is removed. Third feature of person is extracted using Hanavan's Model i.e., model based approach. Finally the person is searched in gait database by using SVM with K-NN and NN .In this research different classifier like SVM, NN and KNN have been used for gait recognition. Pre-processing and then background subtraction enhances feature extraction and recognition process. The results are analyzed and found that accuracy is enhanced when combined classifier SVM,NN and K-NN are used. Experimental Results show accuracy of 98.7% with combined classifiers SVM, NN and K-NN.[14]

VII. PROPOSED WORK

In data mining, clustering is the chief area that has been used in a various applications. Fuzzy C-means clustering is utilized to amass the data into groups by characterizing certain degree. We have used Ant Colony Optimization (ACO) algorithm with Fuzzy C-Means algorithms. ACO is used for solving the problems of computational to make superior results. In our proposed work, we applied to improve the cluster to advance the quality. Crime management is a important and interesting application which has a significant responsibility in society. There was a severe enhancement in this area in the past few years, thus it has become significant for us to create a system for proper crime detection. In today's world, security is a manner which gives higher priority by all political parties and governments worldwide aiming to reduce crime incidence. As data mining is the suitable field to apply for high-volume wrongdoing data collection which learning picked up from data mining methodologies will be a valuable and support police force for crime analysis. So In this paper crime analysis is done by performing Fuzzy C-means clustering algorithm on crime dataset. Thus, we are working in the field of crime detection using an optimized technique of ACO and also improved

algorithm is applied to improve the crime detection capability. Using the above formed, defined flowchart of ant colony optimization, we are going to improve the optimization in the optimization in the field of crime detection.

Proposed Algorithm:

Step:1 Start
Step:2 Apply Fuzzy C-Means to form the clusters
Step:3 Choose cluster centers at random
Step:4 Compute the fuzzy membership
Step:5 Calculate the fuzzy centers
Step:6 Repeat step 4 and 5 until the minimum objective function value is attained
Step:7 Then the output of the above step become the input
Step:8 Now we apply ACO for the detection of the crime at different cities
Step:9 Stop

VIII. RESULT ANALYSIS

We used Fuzzy C means for clustering the data into city wise and year wise. There are certain cities in which we performed the process for the detection of crime.

```
cities =

    'AGRA'
    'BANGALORE'
    'CHENNAI'
    'DELHI'
    'HYDERABAD'
    'JAIPUR'
    'KANPUR'
    'KOLKATA'
    'MUMBAI'
    'PUNE'
```

Fig 1. Name of the Cities

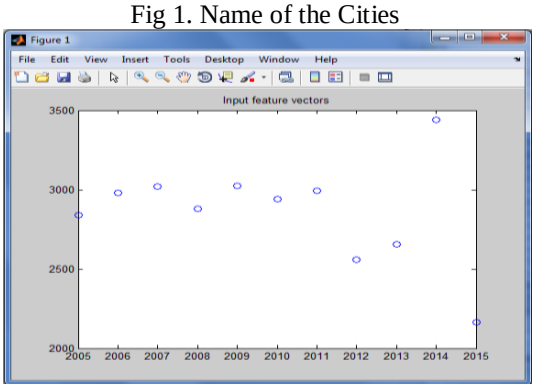

Fig 2. Crime Rate per year

This can be defined as the termination measure at each iteration.

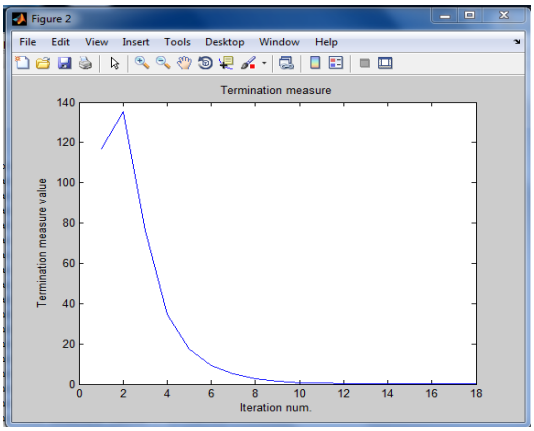

Fig 3. Termination Measure at each iteration

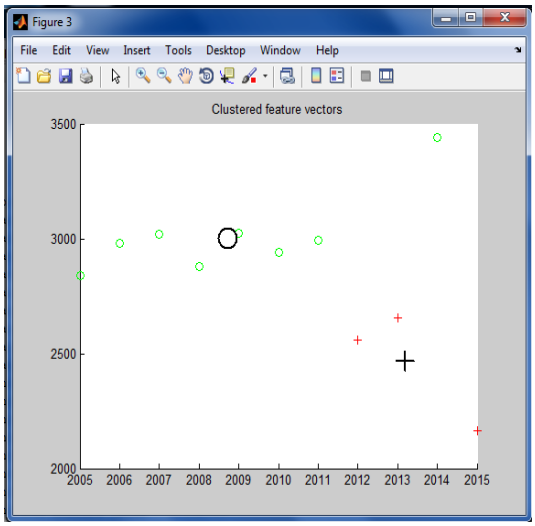

Fig 4. Clusters formation

Membership function used to show the degree of the value to show that proposed work is better.

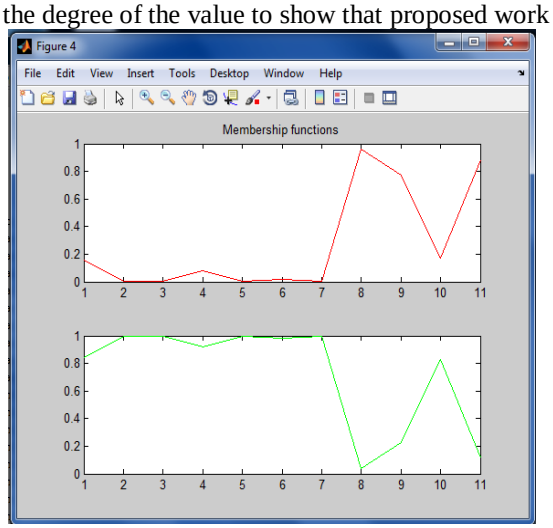

Fig 5. Membership Function

Now we used ACO to show the % increase in crime of different cities. The graph below shows the superior analysis for the different crime count.

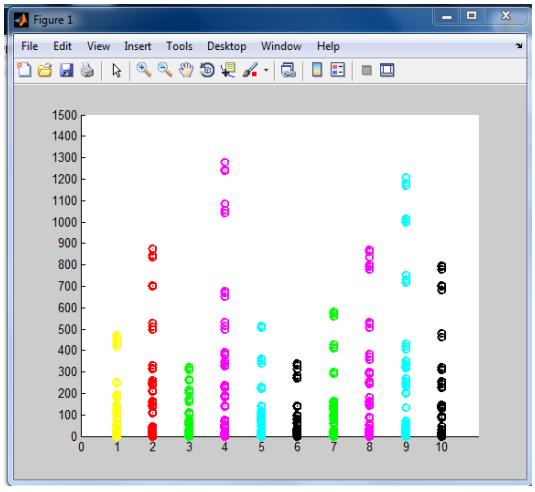

Fig 6. Crime detection using ACO

There we took 5 cities to show the crime rate in different cities in the form of pie chart.

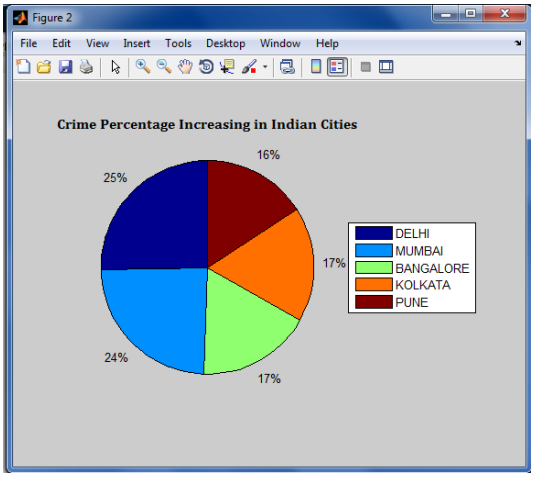

Fig 7. Pie Chart of Crime % increasing in different cities

Conclusion

This paper completes an assortment of methodologies for security conservation data mining and analysis procedures and technique what are exhibited. Crime experts use their knowledge skills, perception and history knowledge when they agreement with criminals and connected crime cases. They are not deviated by the erroneous crime data (e.g., person who is not a criminal but erroneously recognized as criminal by the organization). In this paper, we extract the crime efficiently by our proposed technique and enhance the detection rate. We used different cities to show their crime count in different years. Various techniques used in our proposed methodology by using clustering and ACO algorithms to detect the crime efficiently.

References

[1] Mohammadian, M., "Intelligent Agents for Data Mining and Information Retrieval," Hershey, PA Idea Group Publishing, 2004
[2] Manish Gupta 1 *, B. Chandra1 and M. P. Gupta 1, "Crime Data Mining for Indian Police Information System" ,
[3] Mo Shan , Fei Wang, Feng Lin , Zhi Gao , Ya Z. Tang , Ben M. Chen2, "Google Map Aided Visual Navigation for UAVs in GPS-denied Environment" , 978-1-4673-9675-2/15/$31.00 © 2015 IEEE.
[4] Rasoul Kiani, Siamak Mahdavi, Amin Keshavarzi, "Analysis and Prediction of Crimes by Clustering and Classification", (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.8, 2015.
[5] Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal, "Crime Analysis using K-Means Clustering", International Journal of Computer Applications (0975 – 8887) Volume 83 – No4, December 2013.
[6] Han, J., Kamber, M. and Pei, J., 2011. "Data mining: concepts and techniques". Elsevier publication
[7] J. Han, and M. Kamber, ―Data mining: concepts and techniques,‖ 2nd Edition, Morgan Kaufmann Publisher, 2001.
[8] T.M. Cover, P.E. Hart, "Nearest neighbor pattern classification" , IEEE Transactions on Information Theory IT-13, 21- 27, 1967
[9] Neha Agarwal, Sujeet Singh Bhadouria, "Crime Detection In Rural Areas Using Enhanced Prophet Routing Algorithm in DTN" , 978-1-5090-0669-4/16/$31.00©2016IEEE.
[10] K.Veena1 , P.Visu, "Detection Of Cyber Crime : An Approach Using The Lie Detection Technique And Methods To Solve It" , International Conference On Information Communication And Embedded System(ICICES 2016)
[11]Vukosi Marivate, Pelonomi Moiloa, "Catching Crime: Detection of Public Safety Incidents using Social Media" , 978-1-5090-3335-5/16/$31.00 ©2016 IEEE.
[12]Kalyani C S, Mallikarjuna Swamy M S, "Segmentation of Rectum fromCT Images Using K-Means clustering for the EBRT of Prostate Cancer", 978-1- 5090-4697- 3/16/$31.00 ©2016 IEEE
[13]Abba Suganda Girsang, Fidelson Tanzil, Yogi Udjaja, "Robust Adaptive Genetic K-Means Algorithm Using Greedy Selection for Clustering" 978-1- 5090-5130- 4/16/$31.00 ©2016 IEEE. Taranjot Kaur
[14]Bajwa, Sourav Garg, Kumar Saurabh, "GAIT Analysis for Identification by Using SVM with K-NN and NN Techniques", 978-1- 5090-3669- 1/16/$31.00 ©2016 IEEE.

