


A Tree Based Approach for Data Pre-processing and Pattern Matching for Accident Mapping on Road Networks

Arvind Kumar¹ · Suchi Johari¹ · Deepak Proch¹ · Pardeep Kumar¹  · Durg Singh Chauhan²

Received: 16 April 2015 / Revised: 18 March 2018 / Accepted: 3 April 2018 / Published online: 25 July 2018
© The National Academy of Sciences, India 2018

Abstract Numbers of road accidents are increasing day by day and it is necessary to keep record of all these accidents, for the purpose of providing road safety. The records maintained by the police might have some inaccurate data. For this purpose, different authors have proposed different techniques, in which the accidents are mapped to the correct road segment. The existing accident mapping algorithms have many limitations such as (1) algorithms developed are for the specific datasets only, (2) improper mapping in case of complex road networks and (3) modular approach is used for different types of roads, instead of any general rule to be applied, (4) learning algorithm developed involve lot of cost and efforts, without guaranteeing the mapping of accident to the exact location. The challenge is to identify this inaccuracy in data and correct it with least amount of cost and effort involved. Significance of the attributes having inaccuracy plays an important role in mapping accidents to the correct location. In this paper, our main aim is to (1) map the accidents on the correct location, (2) finding missing values, (3) calculation of the erroneous values in police recorded data and (4) estimation of the attribute significance. The proposed algorithm can map the accidents to correct location instead of mapping an accident to a road segment, junction, and candidate link. The basic concern is that the algorithm being developed should be generalized and can be applied on any type of a road network. To achieve this aim, we have proposed the

minimum bounded region based tree technique, which used for pattern matching methodology.

Keywords Accident-mapping · Pattern-matching · Minimum bounded region (MBR) · Missing values · Attribute significance

1 Introduction

According to National Crime Records Bureau, Ministry of Home affairs, Government of India has announced that in year 2013, a total of 708,478 cases of unnatural accidents have been recorded. Among these accidents 377,758 lives were lost, while 505,368 people were injured [1]. Around 34.3% of the people died in the road accidents. In the year 2012, there were 94.3% of deaths due to unnatural causes among which the role of road accidental deaths was 35.2%. If we consider the daily road accident records according to the report, it was found that there were 377 deaths and 1287 injuries in a day, and 13 people die every hour. It is required that, to control accidents the hot spots where maximum number of accidents occur should be identified properly. The mapping algorithms which can map an accident to a location is required in this situation. But inaccuracies in the police recorded data leads to wrong accident mapping. For the purpose of mapping accidents to the correct location, different parameters are considered which include: country, state, city, road, region, location (in terms of x and y co-ordinates), nearby point and direction of a vehicle. Each attribute has different significance while calculating the inaccuracies in police recorded data. The main problems faced in the records, collected by the police, is the presence of large number of (1) missing values, (2) mistakes in the names of the roads like spelling

✉ Pardeep Kumar
pardeepkumarkhokhar@gmail.com

¹ Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, Solan, H.P., India

² GLA University, Mathura, India

mistakes, (3) same name of different roads, (4) single road can have multiple names. For the enhanced road safety applications, it is mandatory to map the accidents correctly on the road network, so that the hotspots of accidents can be located, where the safety measures can be applied by the government. The main task is to test the correctness of records in the police recorded database based on the road network database. Attributes used for Accident mapping in the proposed approach will include: country, state, city, region, location (in terms of x and y co-ordinates). One more attribute is considered, in case of a two-way lane i.e. direction of vehicle movement on basis of which, we decide the side of two-way lane where the accident has occurred.

2 Related Work

A large amount of inaccurate and missing values are found in the records maintained in the police databases. These records consist of road accident data like road type, road name, type of vehicle, vehicle movement etc. In this section, some existing accident-mapping approaches are discussed, which are used to locate accident to specific location. Authors have been concerning on the safety related to an area analyzed accident records by identifying the location, where accidents occur most frequently [2]. Features like type of vehicle, number of person injured or died are included in GIS database. First, the records of accidents in Kannur district is collected from police station and then, Ground Control Point (GCP) are collected via GPS technology. Spatial and non-spatial data are compared with GIS database and then used to identify location of accident. Author developed a technique, in which clustering of accidents can be cleared from kernel density map, generated through GIS after plotting its location based on accident records [3]. Kernel density map are used for calculating magnitude, area from point feature using kernel function. Author [4] proposed for the two-way associative representing an approach based on spatial relationship between the traffic accident and road network. For reducing accidents as an index, technique needs to extract the hot spots. Researchers [5] has introduced a multilayer model which is based on the concept of artificial neural network (ANN) model for analyzing of the missing values, or prediction and estimation of accidental locations. For categorizing road accidents, clustering techniques with self-organizing map (SOM) was used by author. Author [6] discusses issue in cost sensitive learning that considers both test costs and misclassification costs. If some attributes are too expensive in obtaining their values, it would be more cost-effective to find out their missing values. Author [7] developed iterative approach to predict missing value.

Records in dataset are first partitioned into two groups, one group consists of tuples having no missing values and other group consists of tuples having missing values. Classifier is trained with original dataset, which was used to predict missing values of one group. Feed the data in place of missing entries and process continues until all missing entries get their predicted data. Finally, both groups are combined to form complete dataset. To compute missing values, three techniques are used [8] named as listwise deletion, Mean/mode imputation and KNN imputation. On resulting datasets, C4.5 classification algorithm is applied individually. Author [9] was used k-NN (k-nearest neighbor) for handling missing values in dataset. Author compared k-NN approach with list-wise deletion, mean imputation and proved that k-NN is better than mean and list-wise deletion. Researchers [10] proposed RMS imputation method, which was compared with mean value substitution, random value substitution, and constant value substitution and hence provides better performance. Researchers [11] have introduced imputation, which fills missing values by estimated values on the basis of well-defined procedures. In this method the missing values are replaced by those values which are estimated based on some information. Author [12] has introduced a feature finding mechanism that have high effect to mortality and that of time frame. Author in this paper has suggested the data mining procedures to tackle the complexity of data, missing values, dimensionality reduction, and the problem of estimating or predicting values, and has opted for the methods such replacing missing values, selection of feature, and classification. Researchers [13] have presented an unsupervised learning technique based on a Kohonen self-organizing map used for both categorical and numerical data values with improved accuracy.

3 Proposed Approach

3.1 Minimum Bounded Region/Rectangle (MBR)

MBR is a minimum bounded 2D space containing the object completely, with least amount of the space being left empty. Object being bounded is the location, where the occurrence of accident is recorded in the police records. In Fig. 1, the object is the country India, and the blue color rectangle shows the MBR around India.

Suppose the recorded accident in the police database has records: country = 'India', state = 'Uttarakhand', city = 'Dehradun', region = 'Vasantvihar', location = 2000 sqm area. Our task here is to find MBR for all the recorded data by the police. Figure 2 shows the MBR for this recorded data. To find the minimum bounded region we have used a rectangle, because to bound area, such as a road, rectangle

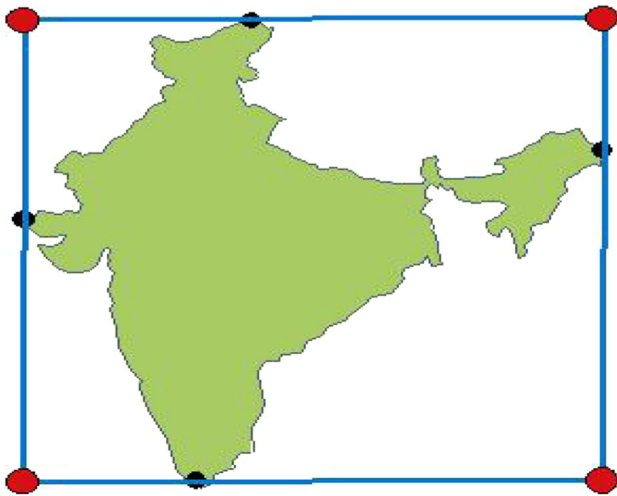


Fig. 1 Minimum bounded region

proves to be a best solution as it requires a least area to bind a segment of road where an accident occurred. MBR for an object can be obtained using the two methods mentioned below:

3.1.1 Method I

To obtain the MBR of a particular region, first of all take four extreme points of the object as: left, right, top, bottom extreme point be L_ext , R_ext , T_ext , B_ext respectively. Consider the x -coordinate value of L_ext be x_{min} and x -coordinate of R_ext as x_{max} . Similarly, the y -coordinate of B_ext is taken as the y_{min} and the y -coordinate of T_ext is taken as y_{max} . With the points x_{min} and x_{max} draw a line perpendicular to y -axis and using points y_{min} and y_{max} draw perpendicular to x -axis. These four lines will intersect each other, and the region obtained by the intersection of these lines is the MBR. Figure 3 shows diagrammatically that how to obtain the MBR for a particular object.

As it is clear from Algorithm 1 that the minimum bounded region is obtained around the object but still it is not the surety that the region obtained is the most optimized. To get least MBR, method II can be used.

3.1.2 Method II

As discussed in the above section, the MBR obtained is along the x -axis and y -axis. But this MBR can be further minimized, if instead of creating the MBR along the x -axis and y -axis, we consider the principal axis. For this purpose, to find the MBR along the principal axis, certain steps can be followed as:

Step I: Calculate centroid of an object For the calculation of the centroid, if the object is regular in shape, then any of the existing methods can be used to compute the centroid mathematically, but if the object is convex i.e. irregular in shape, then the procedure for calculating the centroid would be as discussed in this section. Consider a set of all points lying on the boundary of the object as $P = p_1, p_2, p_3, \dots, p_n$. A single point p_i of the set P is represented in terms of x -axis and y -axis as (x_i, y_i) . For the set P the centroid C is represented as point (C_{x_i}, C_{y_i}) and these points are calculated as:

$$C_{x_i} = \frac{1}{n} \sum_{i=1}^n x_i \quad C_{y_i} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Step II: Obtain principal axis using centroid For the object, set of boundary points are considered again as $P = p_1, p_2, p_3, \dots, p_n$. Let us consider that the angle of the major axis is taken to be α . Now it is considered that the direction of principal axes should be such that the sum of square of perpendicular distances to the principal axis should be minimum [14]. Equation of the line passing through the centroid can be given as:

Algorithm 1: Minimum Bounded Region

Data: Object (area like country, state, city, region etc.)

Result: Minimum bounded region around the object

```

1 Procedure Create_MBR (Region R)
2   Calculate four extreme points of the object as L_ext, R_ext, T_ext, B_ext;
3    $x_{min}$ = $x$ -coordinate of L_ext       $y_{min}$  =  $y$ -coordinate of B_ext;
4    $x_{max}$ = $x$ -coordinate of R_ext       $y_{max}$ =  $y$ -coordinate of T_ext;
5   Draw line passing through  $x_{min}$  in direction of  $y$ -axis, which is perpendicular to  $x$ -
   axis;
6   repeat
7     same for  $x_{max}$ 
8   until get line passing through  $x_{max}$  in direction of  $y$ -axis;
9   Draw line passing through  $y_{min}$  in direction of  $x$ -axis, which is perpendicular to  $y$ -
   axis;
10  repeat
11    same for  $y_{max}$ 
12  until get line passing through  $y_{max}$  in direction of  $x$ -axis;
13  Intersecting points of these lines are taken as vertices;
14  Draw rectangle passing through these vertices;
15  Rectangle obtained is required MBR;
```

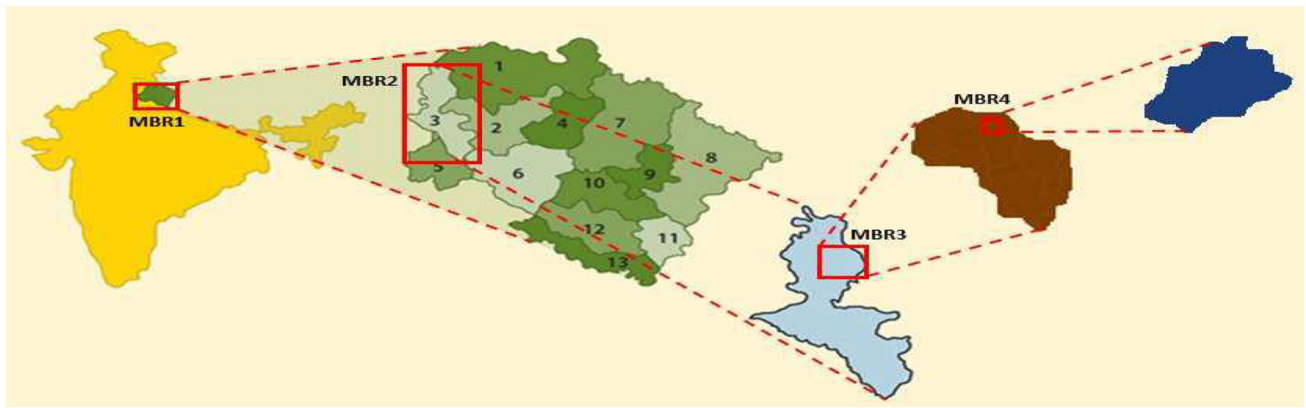


Fig. 2 MBR for accident record as Country = 'India', State = 'Uttarakhand', City = 'Dehradun', region = 'Vasantvihar' and exact location as 2000 sqm

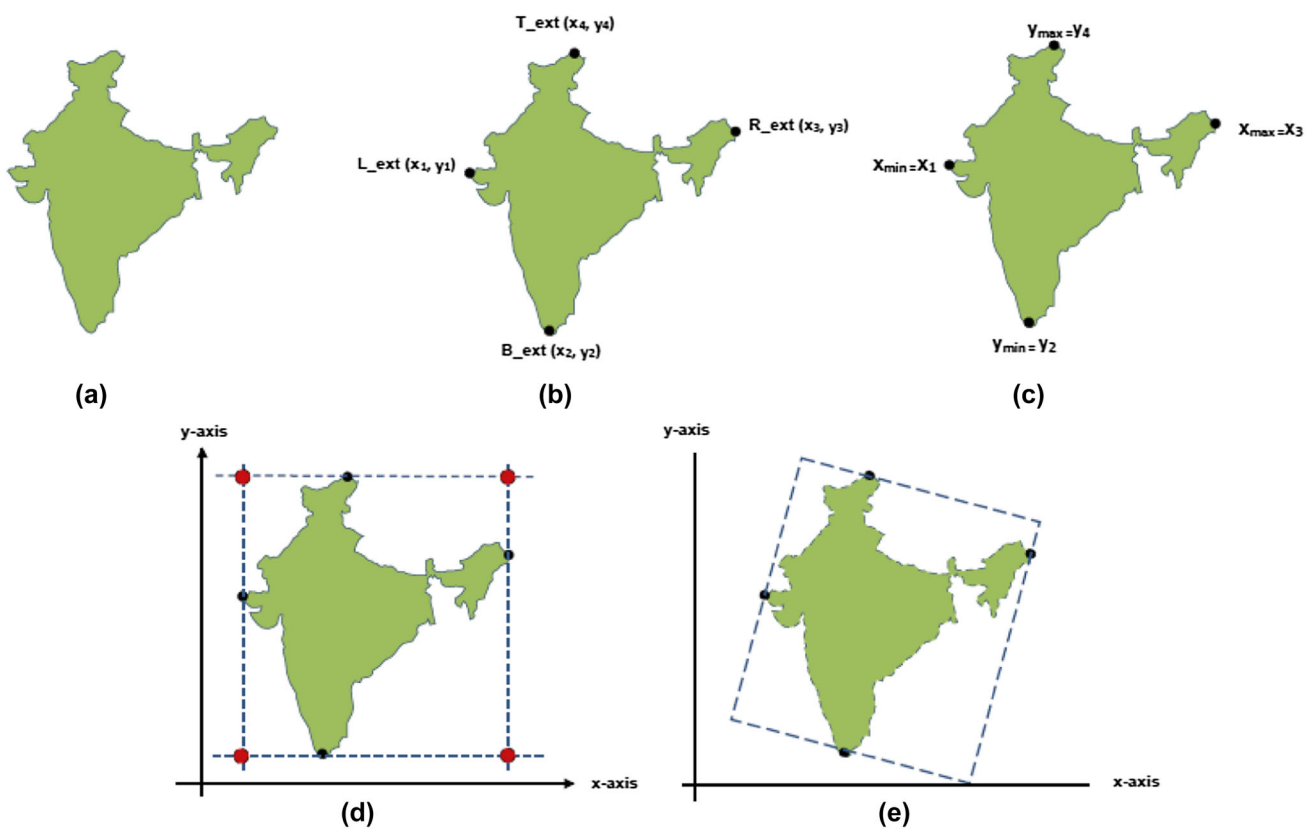


Fig. 3 a Object, b extreme points (top, bottom, left, right), c minimum points in x-axis and y-axis, d MBR using method I, e least MBR using method II

$$x \tan \alpha - y + C_{yi} - C_{xi} \tan \alpha = 0.$$

Perpendicular distance $d_{\perp i}$ of the points in the set $P = p_1, p_2, p_3, \dots, p_n$ can be calculated as:

$$d_{\perp i} = (x_i - C_{xi}) \sin \alpha - (y_i - C_{yi}) \cos \alpha.$$

Now, the sum of square of the perpendicular distance will be calculated as:

$$d_{\perp i}^2 = [(x_i - C_{xi}) \sin \alpha - (y_i - C_{yi}) \cos \alpha]^2.$$

In order to calculate the value of α , differentiation of $d_{\perp i}^2$ has to be done so that the value of $d_{\perp i}^2$ is minimized with respect to α

$$\frac{\Delta d_{\perp i}^2}{\Delta x} = 0 \Rightarrow \tan 2\alpha = \frac{2 \sum_{i=1}^n (x_i - C_{xi})(y_i - C_{yi})}{\sum_{i=1}^n [(x_i - C_{xi})^2 - (y_i - C_{yi})^2]}$$

Since in this formula instead of taking all the internal points of an object all together, the main concentration is on the edge points, and hence this method proves to be better than the other existing methods as lesser the number of edge points in comparison to the internal points hence less calculation involved.

To find the exact position where the point $p(x_a, y_a)$ belongs with respect to major or minor axis the following convention can be used:

$$p(x_a, y_a) = \begin{cases} \text{lies above or right of the line } f(x, y) = 0 & \text{iff } (x_a, y_a) > 0 \\ \text{lies on the line } f(x, y) = 0 & \text{iff } (x_a, y_a) = 0 \\ \text{lies below or left of the line } f(x, y) = 0 & \text{iff } (x_a, y_a) < 0 \end{cases}$$

Step III: Calculate lower and upper edge points Once the principal axis are obtained, then the major concern is to find the points which are at longest distance apart from each other on the principal axis. Let a straight line be represented as a function of x-axis and y-axis as $f(x, y) = 0$.

Equation for major axis:

$$f(x_a - y_a) = (y - C_{yi}) - \tan \alpha(x - C_{xi}) = 0$$

Equation for minor axis:

$$f(x_a - y_a) = (y - C_{yi}) - \cot \alpha(x - C_{xi}) = 0$$

All the points $p(x_a, y_a)$ are upper edge points if $f(x_a, y_a) > 0$ whereas if $f(x_a, y_a) < 0$ then the points $p(x_a, y_a)$ are lower edge points, else the points lies on the axis itself. These upper and lower points are considered in the set $U_{\text{major}}, L_{\text{major}}, U_{\text{minor}}, L_{\text{minor}}$ with respect to major and minor axis respectively. Now the farthest points in the U_{major} and L_{major} are taken as the edge points for the major axis, and the farthest points in the U_{minor} and L_{minor} are considered as the edge points for the minor axis.

Step IV: Compute bounded rectangle vertices for object to be bound The points obtained as the upper edge point and lower edge points for major and minor axis will contribute to obtain the vertices of the rectangle to be drawn as MBR for an object. Let (x_{u1}, y_{u1}) and (x_{l1}, y_{l1}) (be the upper and lower farthest edge points for major axis. The line passing through (x_{u1}, y_{u1}) and which is parallel to major axis can be obtained using Eq. 1

$$(y - y_{u1}) - \tan \alpha(x - x_{u1}) = 0 \tag{1}$$

This line will act as one of the edge of the MBR, in the same way the other edge passing through (x_{l1}, y_{l1}) and parallel to major axis is found using Eq. 2

$$(y - y_{l1}) - \tan \alpha(x - x_{l1}) = 0 \tag{2}$$

For the other two edges the upper and lower furthest edge points (x_{u2}, y_{u2}) and (x_{l2}, y_{l2}) of the minor axis is found, and in the similar way as mentioned above we can obtain the other two lines which would be parallel to minor axis as given in Eqs. 3 and 4

$$(y - y_{u2}) - \cot \alpha(x - x_{u2}) = 0 \tag{3}$$

$$(y - y_{l2}) - \cot \alpha(x - x_{l2}) = 0 \tag{4}$$

Intersection of these four lines gives four vertices required to obtain the least MBR. By solving Eqs. 1 and 3 we are able to obtain the vertices as:

$$v_{tl} = \left(\begin{aligned} tl_x &= \frac{x_{u1} \tan \alpha + x_{u2} \cot \alpha + y_{u2} - y_{u1}}{\tan \alpha + \cot \alpha} \\ tl_y &= \frac{y_{u1} \cot \alpha + y_{u2} \tan \alpha + x_{u2} - x_{u1}}{\tan \alpha + \cot \alpha} \end{aligned} \right)$$

From Eqs. 1 and 4 we obtain:

$$v_{tr} = \left(\begin{aligned} tr_x &= \frac{x_{u1} \tan \alpha + x_{l2} \cot \alpha + y_{l2} - y_{u1}}{\tan \alpha + \cot \alpha} \\ tr_y &= \frac{y_{u1} \cot \alpha + y_{l2} \tan \alpha + x_{l2} - x_{u1}}{\tan \alpha + \cot \alpha} \end{aligned} \right)$$

By solving Eqs. 2 and 3 we obtain:

$$v_{bl} = \left(\begin{aligned} bl_x &= \frac{x_{l1} \tan \alpha + x_{u2} \cot \alpha + y_{u2} - y_{l1}}{\tan \alpha + \cot \alpha} \\ bl_y &= \frac{y_{l1} \cot \alpha + y_{u2} \tan \alpha + x_{u2} - x_{l1}}{\tan \alpha + \cot \alpha} \end{aligned} \right)$$

By solving Eqs. 2 and 4 we get:

$$v_{br} = \left(\begin{aligned} br_x &= \frac{x_{l1} \tan \alpha + x_{l2} \cot \alpha + y_{l2} - y_{l1}}{\tan \alpha + \cot \alpha} \\ br_y &= \frac{y_{l1} \cot \alpha + y_{l2} \tan \alpha + x_{u2} - x_{l1}}{\tan \alpha + \cot \alpha} \end{aligned} \right)$$

The vertices obtained above $v_{tl}, v_{tr}, v_{bl}, v_{br}$ represents the top-left, top-right, bottom-left and bottom-right vertices respectively for the least MBR. So, the minimum bounded region obtained here are most optimized MBR.

3.2 MBR Based Tree

MBR based tree is a tree-based structure in which the hierarchy is maintained. The MBR with the largest area is kept at the root. Leaf nodes consist of the smallest area MBR whose area is 2000 sqm. Internal nodes have the MBR of the area lying between the largest and the smallest area MBR.

Algorithm 2: MBR based tree(MBRT)

Data: Attributes like country, state, city, region, location etc.
Result: MBR based Tree

```

1 While attribute! =null do
2   Procedure Create MBR (Region R)
3   Obtain MBR for region R;
4   Create root node of the tree as the MBR of the country;
5   Leaf node will be the smallest MBR with maximum area of
   2000sqm;
6   Internal nodes consist of the MBR for different parameters
   according to hierarchy such as state, region and location;

```

The hierarchy is maintained from top to bottom according to the area of the MBR. Tree consists of nodes, which represents the MBR for country, state, city, region and location of the accident in the hierarchical order as shown in Fig. 4. Edges of the tree can be left unlabeled as per the requirements. But for the implementation of proposed approach and making the traversal in the tree smoother, we can label the edges with the same label as the

spelling mistakes, for example 'dehradun' spelled as 'dheradun', (3) inaccuracy due to area having more than one name, difficult to decide which name of the location should be noted in the records as multiple names of single place can lead to ambiguity and conflict, for example country India has two more names to refer it as 'Hindustan' and 'Bharat'. With the help of Algorithm 3, pattern matching can be done using MBRT.

Algorithm 3: Pattern Matching using MBR based tree

Data: MBR based tree and police record to be verified
Result: Records verified

```

1 for each police record to be verified do
  Start from root node containing name of the country traverse the
  decision tree using records till leaf node is not reached;
2   if successfully reached leaf node then
3     match found implies that accident mapped correctly, and hence
     police recorded data is correct;
4   else
5     no match found;
6     if (i) missing values in police records, (ii) some wrong entry or some
     mistake is there in the database or (iii) more than one names for the
     same location, or (iv) different locations can have same name.

```

node to which the edge is pointing. Algorithm 2 shows how to create MBR based tree.

3.3 Algorithm for Pattern Matching Using MBR Based Tree

In the MBR based tree, our main task is the pattern matching which help us to obtain the region where accident has occurred. This is the smallest region of 2000 sqm which has to be found. From the police records, different parameters are obtained, whose values are searched in the MBR based tree. While traversing in the tree, if leaf node is reached, then it implies that the accident is correctly mapped, and hence concluded that the accident location is correctly recorded by the police. In case the leaf node is not reached, then we can consider that there was inaccuracy in the police recorded data. Inaccuracy can be of many types among which we are dealing with three major issues, which includes (1) missing values, when there are some fields in the records, left blank by the police, (2) inaccuracy due to

For the purpose of finding the inaccuracies in the police recorded data, we have proposed the technique of MBR based tree (MBRT). This tree consists of the data from the real road network which is accurate. We choose records from police recorded database and traverse in the tree using the entries of the record, such that if successfully reached the leaf node, means match is found i.e. accident mapped correctly, and hence records are correct otherwise not. In this mechanism, database which consists of tables having attributes such as the country, state, city, region, and x-y coordinates determining the location of the accidents as according to the police records. Along with this information, we have some more attributes which consists of road name, road type, vehicles direction and co-habitat (near-by location). Country, state, city, region and location are the parameters to determine the location of the accident on the network, whereas road type and vehicle direction are the two other attributes which contribute to the accuracy of the determination of the accident location. Different types of roads include

one-way lane, two-way lane, slip road, turn-around road and junction. If accident has occurred at the two-way lane, then direction of vehicle plays an important role in determination of the exact location of the accident by determining the lane on which the vehicle was moving. First, MBR of the country is found in the MBR based tree, then moving down hierarchically in the tree, MBR of the state is found after this the MBR of region and location are determined. MBR of the final location of the accident should not be more than a square (special type of rectangle) of area 2000 sqm, which is represented as the leaf of the MBR based tree. Internal nodes consist of the state, region and location according to hierarchy with the maximum size MBR at the top to the minimum at bottom. The MBR based tree is not very complex as the height of the tree is directly proportional to the number of attributes used to determine the location of the accident on the road network. It is necessary to determine the path from the root to the leaf node, if the path exists, then it is considered that the pattern matching is successfully completed, and accident mapped properly, determining the recorded data to be correct. Otherwise no match is found, means there is some kind of inaccuracy in records and hence it is necessary to determine the existence of inaccuracy in the data. The inaccuracy can be due to missing value or erroneous value.

3.4 Calculation of Missing Values Using Proposed Approach

Suppose we are at some node at level 'k' in the decision tree. We are not able to move further, as there is no child node at level 'k + 1' having the same value as being recorded in the police database. So, there might be some error in the records. Let us assume that there is a missing value, so to obtain the correct records let us check the next level 'k + 2' for all the nodes in level 'k + 1', whose parent is the already traversed node of level 'k'. If the value exists in the level 'k + 2', then we can conclude that there was a missing value and the value can be obtained from the level 'k + 1' corresponding to which the value is found at level 'k + 2'.

Figure 5 shows that there is a missing value for the city in the police recorded data. The missing value is shown in the diagram with a question mark. This level is considered as level 'k + 1' and its parent level considered to be the level 'k' with child level of missing level value as level 'k + 2'. To calculate the missing value, take all the child levels of level 'k' and for each one of them search for the match in the next level. If match is found, then it is concluded that the value in level 'k + 1' for which the match was found is the correct value to be filled in the missing value column. For the purpose of finding missing using proposed approach, we use Algorithm 4.

Algorithm 4: Correction of Missing Values

Data: Record with missing value

Result: Correct value to fill missing value

```

1 Traverse the tree from root to leaves;
2 if reached leaf node then
3   | No missing value;
4 else
5   | Missing value is there at Level k;
6   | if missing value found then
7     | for all childs at level 'k+1' for level k do
8     |   | Search for the value at level k+2;
9     |   | if value found then
10    |   |   | The value at level k+1 is required missing value;
11    |   |   | break;
12    |   | else
13    |   |   | consider level k+1 as k, k+2 as k+1 and the child level
14    |   |   | below the original level k+2 as k+2;
15 repeat
16 until till the missing value is not obtained or either the leaf nodes are
    reached;
```

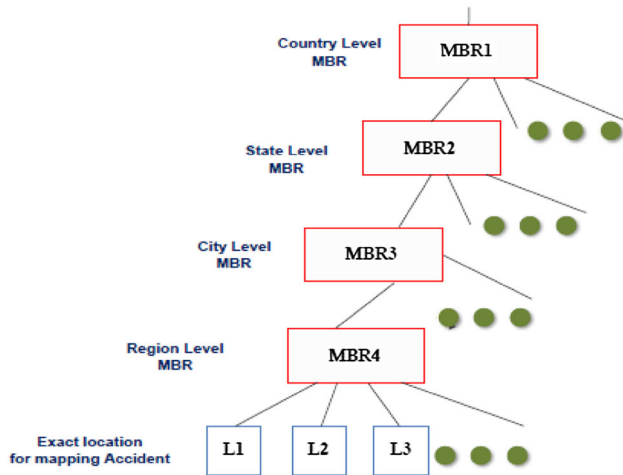


Fig. 4 Minimum bounded region based tree (MBRT)

If there are two missing values at consecutive levels, then the same process of finding the missing value is called recursively. Value of levels will be updated in each iteration by keeping the level with missing value as level ‘ $k + 1$ ’ in every iteration. This procedure is shown in Fig. 6 where first iteration is shown with blue color and second iteration with brown color. Missing value 1 is considered to be at level ‘ $k + 1$ ’ in first iteration, whereas missing value 2 is considered to be at level ‘ $k + 1$ ’ in 2nd iteration and so on. Using this approach, it is possible to find any number of missing values, but accuracy will be affected in case of multiple missing values. In this no extra effort of learning are involved, hence easier to implement and less cost is involved. The proposed algorithm helps not

only to map the accidents to correct location but also to find missing values. It is more efficient to use machine learning algorithm along with the proposed algorithm to find missing values, when there are a large number of missing values in the accident records.

The second case may arise, when there is more than one missing value and that too are not consecutive, as shown in Fig. 7. Different iterations are considered for different missing values and for every iteration the missing value level is considered to be level ‘ $k + 1$ ’. Missing value, once obtained, can be replaced in the record and it can be observed that the correct mapping of the accident on the road network is obtained. Machine learning algorithm can also be used in case there are large numbers of missing values.

3.5 Calculation of Wrong Values Using Proposed Approach

Let there exist a wrong entry in the recorded data of the police, such as spelling mistake, or a local name of the accident location. The spelling mistake can also be resolved using the proposed technique, following the same procedure as calculation of missing value. Assume that the values at level ‘ $k + 1$ ’ are wrongly recorded, and then the node at the level ‘ k ’ is checked. All the values at the level ‘ $k + 1$ ’ which are child of node at level ‘ k ’ are checked for the corresponding recorded value, but if the value is not present then the next recorded attribute value is checked at the level ‘ $k + 2$ ’, if the value is found at level ‘ $k + 2$ ’, then the value at level ‘ $k + 1$ ’ corresponding to the value

Fig. 5 Single missing value

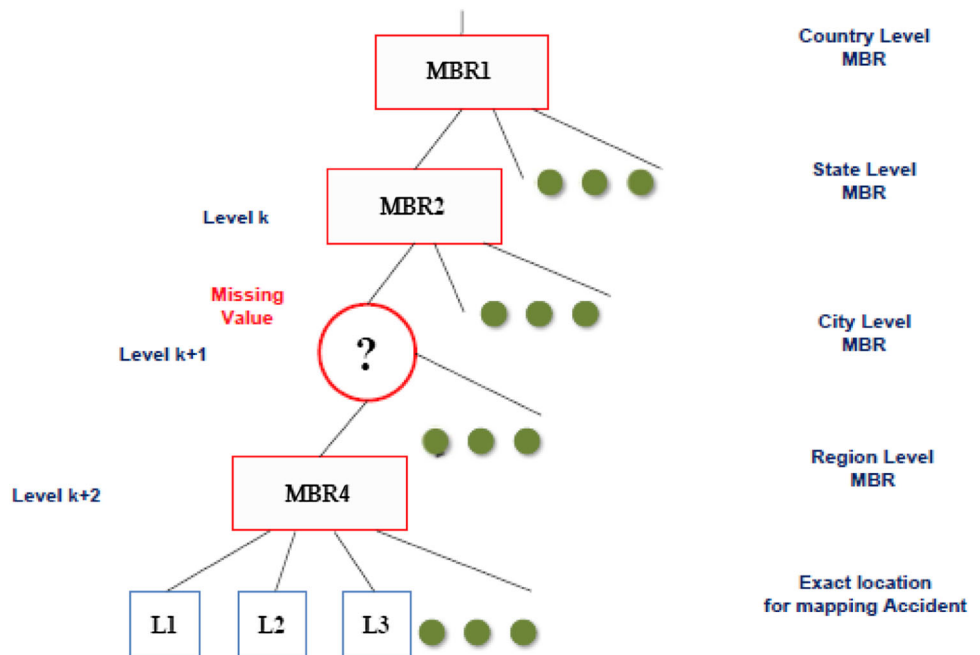


Fig. 6 More than one missing values (consecutive)

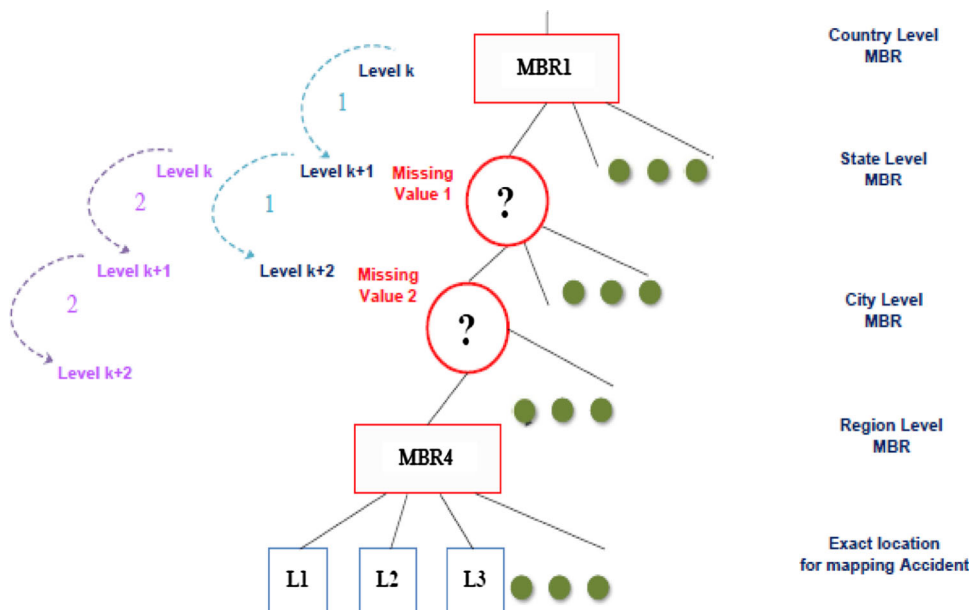
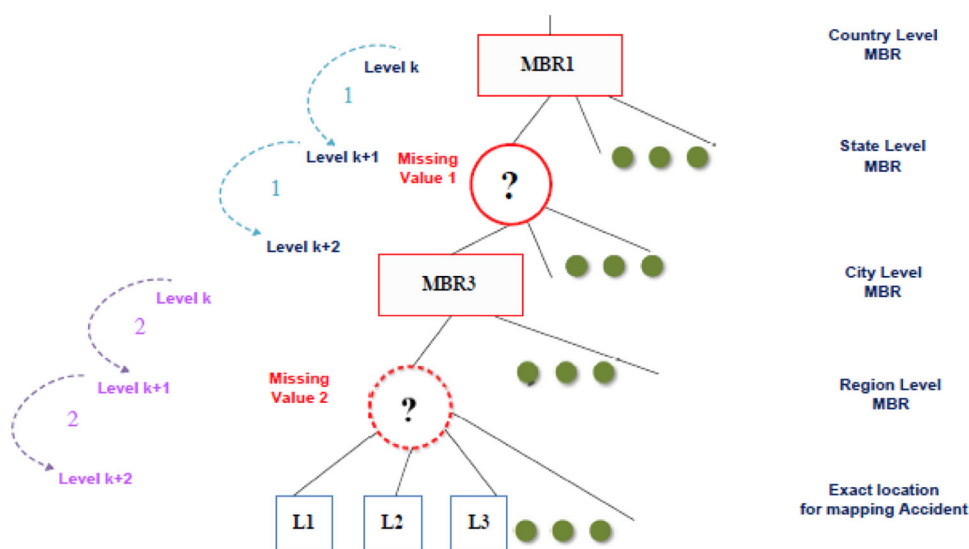


Fig. 7 More than one missing values (non-consecutive)



at level 'k + 2' is taken as the correct value for the wrong records.

important to be recorded correctly by the police in the records.

3.6 Estimation of Attribute Significance in Accident Mapping

Significance of the attribute is defined as, the importance of the attribute to map the accident correctly on the MBR based tree. If the MBR based tree is considered, then the nodes higher in the hierarchy are more important to be recorded correctly. So, the significance of the nodes moving down the tree decreases with the root having highest significance while the leaves having the least significance. It is more difficult to analyse the missing values or erroneous values at the higher levels, so they are more

4 Methodology

MBRT is a technique, which is used not only to map the accidents to their particular location, but also to find the missing and erroneous values and replace them with the correct solution. Figure 8 shows the detailed explanation of the proposed methodology in the form of a flowchart. The methodology used for the MBRT based accident mapping can be divided into certain steps which include:

Step 1 Traverse MBRT and map police recorded data on MBRT.

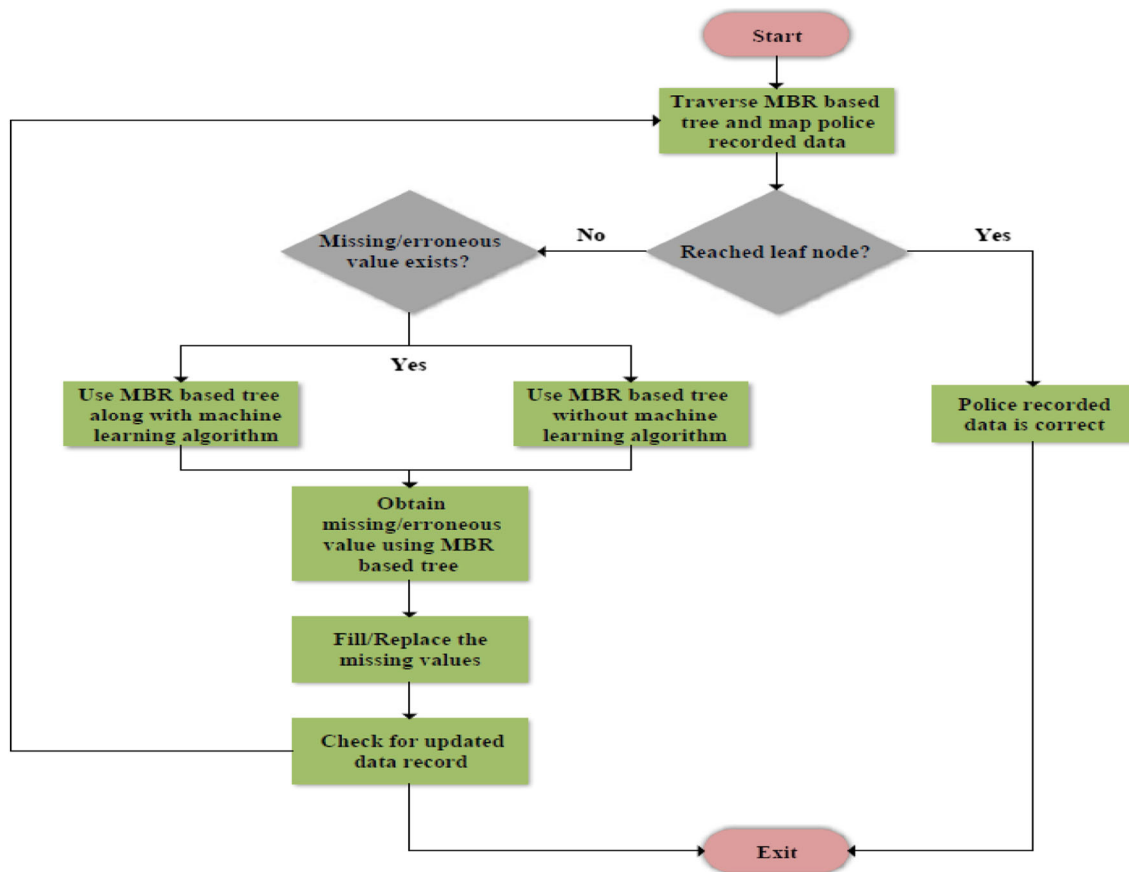


Fig. 8 Flowchart of proposed methodology

Step 2 If there are missing/erroneous, then there are two ways to solve this issue, either use MBRT without any machine learning algorithm involved as in step 4, or MBRT with machine learning algorithm as in step 3.

Step 3 Use MBRT without machine learning Algorithm in Sect. 3.4, can find the solution to the missing or erroneous values.

Step 4 Use MBRT with machine learning algorithm By providing some intelligence using machine learning algorithm such as, ANN leads to improvement in the accuracy and speed.

Step 5 Use MBRT to obtain missing/erroneous value.

Step 6: Fill/replace the missing/erroneous value Correct value obtained in the above step is filled.

Step 7 Check for updated data: Once the wrong record is updated, then check for the updated data and go back to step 1.

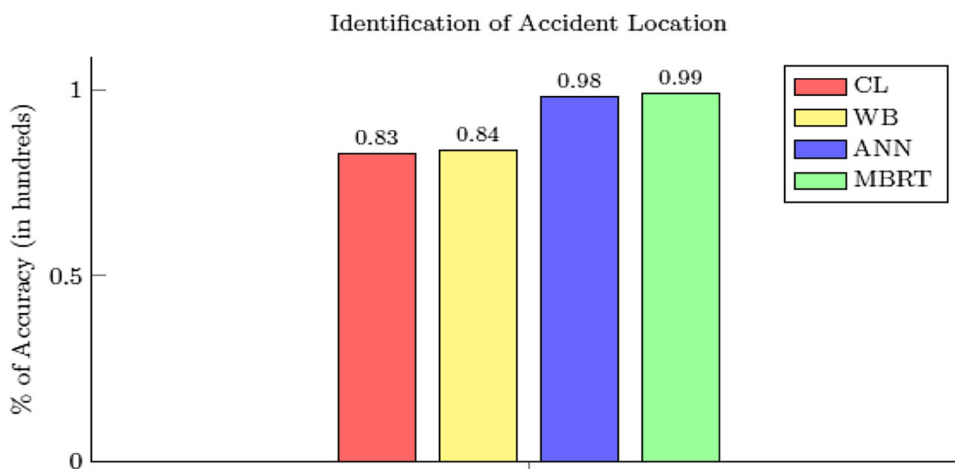
Step 8 Recorded data is correct, and successfully mapped on the MBRT, so exit.

5 Experimental Results

We have simulated the proposed approach on window 8 operating system with 4 GB RAM, NetBeans 8.0 IDE with JDK 8 51, road accident dataset along with JMapView library. Dataset consist of five attributes of the police recorded data, with more than hundred accidental records. Police recorded attributes (categorical attributes) include country, state, city, region and location of the accident. In this section the experimental results obtained are discussed. In Fig. 9, the comparison between the various techniques: closest-link, Weight-based, Artificial Neural Network (ANN), MBR based tree (MBRT), used to map accidents to nearby junction, and candidate link are given. The performance of ANN based algorithm has similar effect as the proposed algorithm because MBRT and ANN can identify the nearby junction and candidate link correctly. Result varies, if exact location of the accidents has to be found instead of the candidate link or nearby junction.

Figure 10 shows that how accurately the different techniques can map the accident correctly to the exact location. Closest-link and Weight-based algorithms could not identify the exact location of the accident, while ANN

Fig. 9 Accuracy in the identification of nearby junction, candidate link for accident mapping



due to some predefined learning procedure can find the exact location to some extent, otherwise ANN also concentrates to find closest candidate link, or nearby junction [15]. MBRT can identify the exact location of the accident in the specified rectangular area of 2000 sqm, and hence

proved to be 85–95% more accurate as compared to other accident mapping techniques.

Off-road accidents are more difficult to map as compared to the other accidents. In case the accident mapping of the off-road accidents has to be done, then it is found

Fig. 10 Accuracy in the identification of exact location of the accident for accident mapping

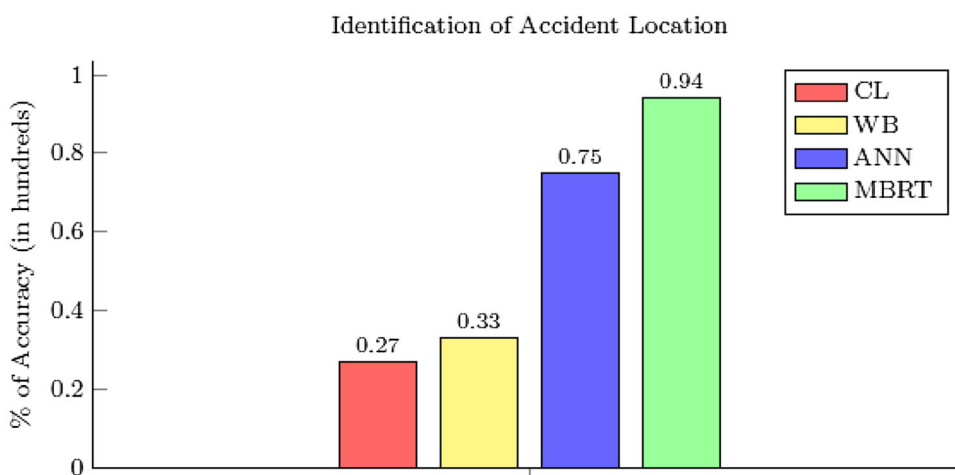


Fig. 11 Accuracy in accident mapping for off-road accidents (machine learning algorithm used)

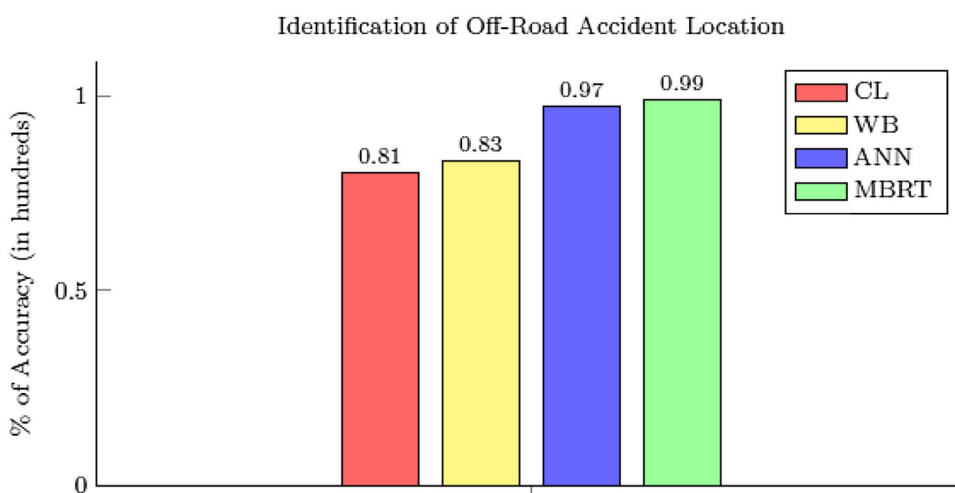


Fig. 12 Accuracy in accident mapping for off-road accidents if MBRT doesn't make use of machine learning

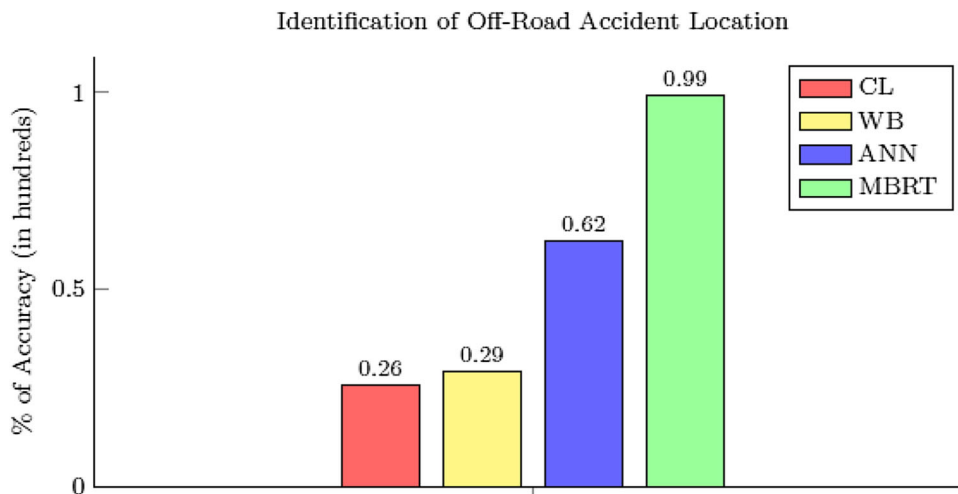
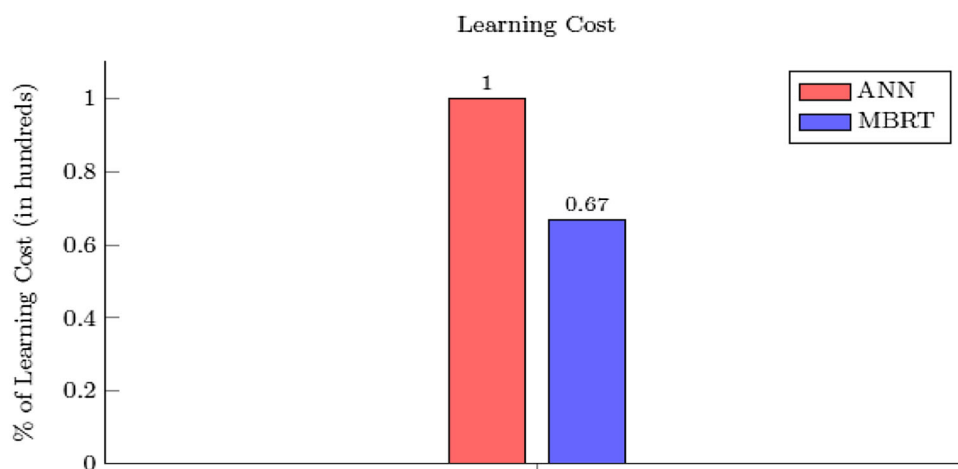


Fig. 13 Machine learning cost required in case of calculation of the missing values (if MBRT doesn't use machine learning algorithm)



that MBRT is equivalent to ANN due to its performance as shown in Fig. 11 but MBRT proves to be better than ANN as no learning cost is involved in MBRT.

Mapping of Off-Road accidents to the exact location is even more difficult. Since the proposed approach concentrates on finding the MBR to locate the accident, so off-road accidents are also treated as similar to other accidents, and hence can be mapped to exact location accurately. No modular approach is required here to map different type of accidents, as in other approaches. Comparison between different techniques is shown in Fig. 12. If we compare the proposed approach with the ANN on the basis of the learning cost involved in the two approaches we conclude that ANN involves around 90–100% of learning cost whereas the proposed algorithm will consider only 50–60% of the learning cost as shown in Fig. 13.

In proposed approach for the mapping of the accident to the correct location there is no involvement of the learning cost as shown in Fig. 14. Learning cost is involved only when some learned algorithm is applied to search for large number of missing values.

Figure 15 shows the comparison between the different techniques to find missing values. The accidental data set is taken and then some of the values were left blank. After this the effect of different algorithms such as mean, mode, k-nearest neighbor, artificial neural networks and MBR based tree are analyzed for different number of missing values on the same data set. Mean and mode algorithm produce same results, similarly the results produced by ANN and MBRT are nearly equivalent and even more accurate than other approaches.

6 Improvements Over Previous Approach

In earlier approaches like closest link, weight-based and ANN the specific location of data could not be found, but with MBRT, accident can be mapped to the specific location with the bounded square of area 2000 sqm. The approaches used till now are modular and perform differently for different types of road network, but MBRT is a generalized form and can be applied on any type of a road

Fig. 14 Machine learning cost required in case of calculation of the missing values (if MBRT uses machine learning algorithm)

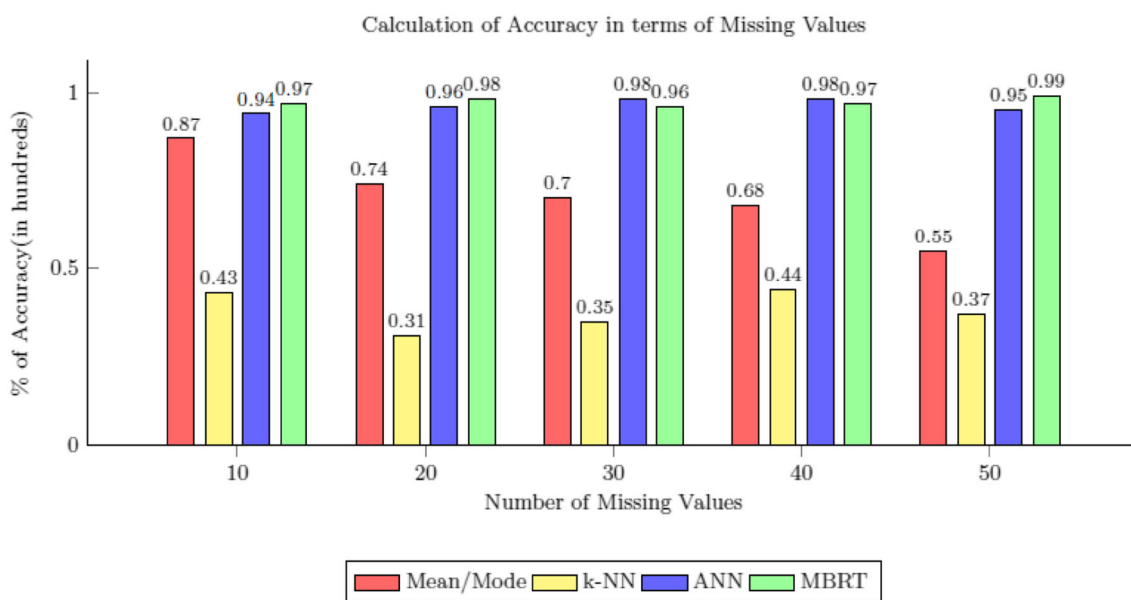
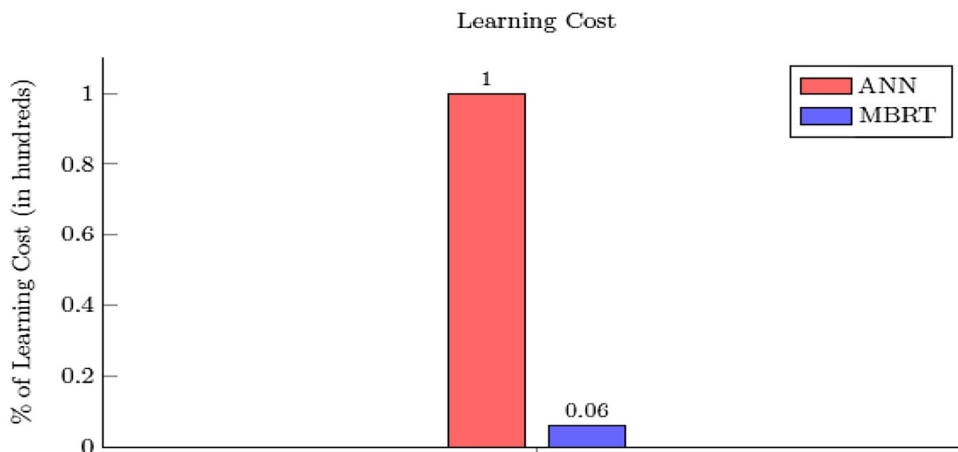


Fig. 15 Comparison of accuracy in calculation of missing values using different approaches

network. In MBRT, learning cost is negligible, required only when large number of missing values. Run-off road accidents were difficult to map using existing approaches, but MBRT can map the run-off accidents correctly to exact location. Table 1 shows the accuracy of mapping and

identification of location where accident has occurred using different techniques.

If we compare the proposed approach with the existing techniques to find the missing value, then we conclude that it produces better results as compared to mean, mode,

Table 1 Accuracy of mapping and identification of accident location using different approaches

	Closest-link (%)	Weight-based (%)	Artificial neural network (%)	MBR based tree (%)
1. Accuracy in identification of accident location (nearby junction or candidate link)	83	83.70	98.40	98.70
2. Accuracy in identification of accident location (exact location)	27	33.20	75.80	94.20
3. Off road accident mapping with machine learning algorithm used in MBRT	80.50	83.20	97.30	98.70
4. Off road accident mapping without machine learning algorithm used in MBRT	25.70	28.92	62.30	98.70

Table 2 Accuracy in calculation of Missing values using different approaches

Missing values	Mean (%)	Mode (%)	k-NN (%)	ANN (%)	MBRT (%)
10	87	87	43	94.20	97
20	74.47	74.47	30.70	96.40	98.20
30	69.70	69.70	34.90	98.90	97.80
40	68.80	68.80	43.47	98	97.70
50	54.90	54.90	37.22	95.37	98.98

k-NN methods, and it produces comparable results to ANN as mentioned in Table 2. In the Table 2, the values 10, 20, ..., 50 represents the number of missing values.

7 Conclusion

Our proposed approach MBRT proves to be better than previous techniques such as closest link or weight-based search, as this algorithm is intelligent enough to fill the missing values and correct the erroneous data. If the proposed approach is compared with the existing ANN based algorithm, then MBRT has reduced the amount of extra effort involved in learning. MBRT also proved to be the best solution to map accidents correctly to exact locations. Our approach is decision tree based approach which discovers the significance of attributes for accident mapping using information gain and the level of significance is as follows: country name (highest significance level) to region (least significance level) for accident mapping.

References

- National Crime Records Bureau (NCRB) (2013) Accidental deaths and suicides in India Report <http://ncrb.gov.in/StatPublications/ADSI/ADSI2013/ADSI-2013.pdf>
- Jayan KD, Ganeshkumar B (2010) Identification of accident hot spots: a GIS based implementation for Kannur district, Kerala. *Int J Geomat Geosci* 1(1):51–59
- Leon DRM, Doroy N, Lidasan H, Castro J (2013) Black spot cluster analysis of motorcycle accidents. In: Proceedings of the Eastern Asia society for transportation studies, vol 9, pp P369 (1–14)
- Chen H (2012) Black spot determination of traffic accident locations and its spatial association characteristic analysis based on GIS. *J Geograph Inf Syst* 4(6):608–617
- Ogwueleka NF, Misra S, Chibueze O, Sanz FL (2014) An artificial neural network model for road accident prediction: a case study of a developing country. *Acta Polytech Hung* 11(5):177–197
- Zhang S, Qin Z, Ling XC, Sheng S (2005) Missing is useful: missing values in cost-sensitive decision trees. *IEEE Trans Knowl Data Eng* 17(12):1689–1693
- Rahman MM, Davis ND (2013) Machine learning based missing value imputation method for clinical dataset. *IAENG Trans Eng Technol* 229:245–257
- Sharma A, Mehta N, Sharma I (2013) Reasoning with missing values in multi attribute datasets. *Int J Adv Res Comput Sci Softw Eng* 3(5):1035–1043
- Minakshi Vohra R, Gimpy M (2014) Missing value imputation in multi attribute data set. *Int J Comput Sci Inf Technol* 5(4):5315–5321
- Somasundaram SR, Nedunchezian R (2012) Missing value imputation using refined mean substitution. *Int J Comput Sci Issues* 9(4):306–313
- Mehala B, Thangaiah JRP, Vivekanandan K (2009) Selecting scalable algorithms to deal with missing values. *Int J Recent Trends Eng* 1(2):80–83
- Poolsawad N, Kambhampati C, Cleland FGJ (2011) Feature selection approaches with missing values handling for data mining—a case study of heart failure dataset. *Int J Biomed Biol Eng* 5(12):531–540
- Singh N, Chabra S, Javeed A, Kumar P (2015) Missing value imputation with unsupervised kohonen self organizing map. *International conference on emerging research in computing, information, communication and applications*. Springer, New Delhi, pp 61–76
- Chaudhuri D, Chaudhuri BB (1991) Elliptic and circular fit of a two-tone image and their 3-D extensions. *Assoc Adv Model Simul Tech Enterp Rev* 15(4):15–36
- Deka L, Quddus M (2014) Network-level accident-mapping: distance based pattern matching using artificial neural network. *Accid Anal Prev* 65:105–113