# Spatial Decision Tree for Accident Data Analysis

J. M. Manasa[1], Shrutilipi Bhattacharjee[2], Soumya K. Ghosh[3], and Sudeshna Mitra[4]

School of Information Technology[123], Department of Civil Engineering[4]
Indian Institute of Technology Kharagpur
West Bengal, India, 721302
Email:manasa.jm91@gmail.com[1], shrutilipi.2007@gmail.com[2], skg@iitkgp.ac.in[3], sudeshna@civil.iitkgp.ernet.in[4]

***Abstract*—Accident data analysis deals with identifying a set of conditions of accident occurrences and the importance of the corresponding implication. It is of prime importance because it gives an insight into the reasons behind the number of fatal and other major injuries. Accident data has an inherent spatial context associated with it, as the location of the accident has an important role to play in its severity. This paper aims at categorizing and analyzing the accident data and drawing some meaningful inferences, that are implicit to the data. A *spatial decision tree* based approach has been used and implemented to draw some useful conclusions, which are spatially relevant with the severity of the accident. The experimentation has been carried out on the accident dataset, collected from the National Highway (NH6) connecting Kharagpur and Kolkata, India. The results exhibit some latent patterns, useful for further accident management.**

***Keywords*—*Accident Data Analysis, Spatial Data Mining, Spatial Decision Tree.***

## I. Introduction

Spatial data mining deals with the discovery of meaningful, coherent and currently unknown inferences from large spatial datasets. The major difference between spatial data mining (SDM) and classical data mining (CDM) lies in the inherent difficulty in handling spatial data, complexity of the data types involved, and the underlying spatial relationships. SDM has drawn a significant research interest owing to its large scope, and the voluminous amount of available spatial datasets. It is the most useful choice to capture new patterns, and dependencies that are present in these data sets.
*Decision trees* can be described as a top-down, divide-and-conquer classification strategy based on automatically selected rules, that partition a set of given entries into smaller classes [1] [2]. SDM leverages on the inherent spatial information underlying geographic datasets and helps in drawing useful inferences from them.
This paper aims at inferring some patterns from the real time accident data, collected from the National Highway between Kharagpur and Kolkata (NH6), India, using the *spatial decision tree*. It contains information such as the time of accident, weather conditions, side of the road, road condition, nature of accident, location and type of accident. Accident data have a very important spatial context and that is leveraged by the *spatial decision trees* to extract new inferences. The aim is to detect patterns over the temporal scale and consequently locate areas which are prone to accidents. This knowledge is captured by *spatial decision trees* and it deals with the location content to provide some spatially significant results.

Many literatures have reported the usefulness of different pattern analysis approaches for accident data analysis. Chang *et al.* [3] have used classification and regression trees to analyze the frequency of freeway accidents. Chang *et al.* [4] have also applied the same strategy to extract the relationship between injury severity and other environmental conditions. Ross Quinlan's have reported the most popular *decision trees* algorithm, ID3 [2], to generate the smallest DT. An extension of the decision trees was proposed by Ester *et al.* [5] to deal with the spatial databases based on ID3. The *decision tree* algorithm for machine learning, integrated with Geographic Information Systems (GIS), was first introduced by Zhang *et al.* [6]. Sitanggang *et al.* [7] have proposed a new *spatial decision tree* algorithm for different discrete elements represented by points, lines and polygons. A new method called the *spatial classification and regression trees* (SCART) was proposed by Chelghoum *et al.* [8] for the classification of spatial data. Ghimire *et al.* [9] have talked about spatial autocorrelation and the use of *decision trees* in accident data analysis. They have shown results supporting the fact that spatial entropy is less when compared to the conventional entropy in classification with respect to accident data analysis.

This paper has been divided into the following sections. Section II describes the theory of *spatial* and *classical decision trees*. The overall flow of the process and an algorithmic representation of the process of drawing spatial inferences is described in Section III. Section IV presents the experimentation and analysis of the results. Finally, the conclusion is drawn in Section V.

## II. Decision Trees

In the field of *decision tree* learning, Iterative Dichotomiser 3 (ID3) is one of the most widely used algorithms, proposed by Ross Quinlan [2]. Every *decision tree* algorithm involves data which has a target attribute and some supporting attributes. The *decision tree* algorithm revolves around choosing the best attribute in each iteration, to further classify the data. ID3 works as a top-down, greedy search to choose the best tree node attribute at each split, among all the supporting attributes. This choice is made on the basis of the criterion called *Entropy*, and the *information gain*. The selection of the tree node is made with respect to these two terms. The supporting attribute with the maximum change in entropy and correspondingly yielding in maximum information gain is chosen as the tree

node. If there are $n$ number of categories in the domain of the target attribute TA, i.e., if the target attribute TA can acquire $n$ different values, *Entropy* can be defined as,

$$Entropy(TA) = -\sum_{i=1}^{n} \mathrm{P}_i \log_2 \mathrm{P}_i$$

Where $P_i$ is the ratio of the number of elements belonging to the $i^{th}$ category to the total number of records. Let SA be a supporting attribute, and $TA_v$ denote the subset of the tuples for which the corresponding value of the supporting attribute is $v$ for each record. $|\mathrm{TA}_v|$ and $|\mathrm{TA}|$ denote the cardinality or the number of elements present in the sets $TA_v$ and TA respectively. The information gain with respect to each supporting attribute is defined as the reduction in entropy, which occurs while partitioning the domain set according to the values of the supporting attribute. It is given as,

$$Gain(TA, SA) = Entropy(TA) - \sum_{v \epsilon Values(\mathrm{SA})} \frac{|\mathrm{TA}_v|}{|\mathrm{TA}|} \mathrm{Entropy}(\mathrm{TA}_v)$$

The tree is built by selecting the best supporting attribute at each level and partitioning the consequent tuples accordingly. A spatial *decision tree* following the ID3 algorithm is described in Section II-A.

### A. Spatial Decision Trees

A *spatial decision tree* can be formed when the spatial aspects and dimensions of the data are leveraged and incorporated within the *classical decision tree*. Entropy in *spatial decision tree* is denoted by $Entropy_s$. The measures for spatial diversity were proposed by Claramunt *et al.* in [10]. The first measure is the intra-distance $d_i$, which can be defined as the average distance between the entities of a same class *i*. The second measure is termed as the extra distance, ${d_i}^{ext}$ of a given class *i*, which encompasses the average distance between the entities of that particular class, with those of the other classes. They can also be written as:

$$\begin{aligned}
{d_i}^{int} &= \frac{1}{|C_i| * (|C_i| - 1)} \sum_{j \epsilon C_i} \sum_{\substack{k \epsilon C_i \\ k \neq j}} \text{dist(j,k) if } |C_i| > 1 \\
{d_i}^{int} &= \lambda, \text{otherwise} \\
{d_i}^{ext} &= \frac{1}{|C_i| * |C - C_i|} \sum_{j \epsilon C_i} \sum_{k \epsilon C - C_i} \text{dist(j,k) if } C_i \neq C \\
{d_i}^{ext} &= \beta, \text{otherwise}
\end{aligned}$$

C represents the set of spatial entities of a given dataset, $C_i$ can be defines as the subset of C belonging to the $i^{th}$ category of the classification, ${d_i}^{int}$ denotes the average distance between the entities of $C_i$, and correspondingly ${d_i}^{ext}$ is the distance between the entities of $C_i$ and the entities of other categories. The distance between two entities *j* and *k* is defined by dist(j, k). To avoid null value exceptions and to prune noise, a relatively small constant $\lambda$, and another relatively high constant $\beta$ are taken. Changes in the spatial distribution of the dataset might change the spatial entropy, owing to the way the spatial entities are distributed in space, even when the category constitution remains same. The spatial entropy measure $Entropy_s$ is achieved by using the intra and extra distances between the spatial objects and incorporating them in Entropy(TA). Further, the spatial information gain is given by,

$$Entropy_s(TA) - \sum_{v \epsilon Values(\mathrm{SA})} \frac{|\mathrm{TA}_v|}{|\mathrm{TA}|} Entropy_s(\mathrm{TA}_v)$$

## III. Spatial Decision Tree based Accident Data Analysis

The proposed framework deals with extracting some meaningful inferences from the *decision trees* computed on the dataset. Some important results can be inferred from these *decision trees*, for example, spotting the hot spots for accidents and investigating the reasons leading to them. The *classical decision trees* could not produce the above conclusion as they do not take the spatial aspect or the location information into account. In order to incorporate the spatial context, the data is spatially sorted on the basis of the distance between a fixed point and the accident location. Two nearby tuples in the spatially sorted database contains information about accidents taking place in nearby locations. It is a general observation that the tuples which are spatially near, are more clustered in the *spatial decision tree*, than those in the *classical decision tree*. Thus, it can be stated that, in a particular area, the accidents occurred due to similar factors and in similar types of situations.

To infer various implicit patterns from the *spatial decision trees*, two parameters are used, namely *support* and *confidence*. The support and confidence for each node can be defined as,

$$Support = \frac{p_i}{N}$$

$$Confidence = \frac{n_i}{p_i}$$

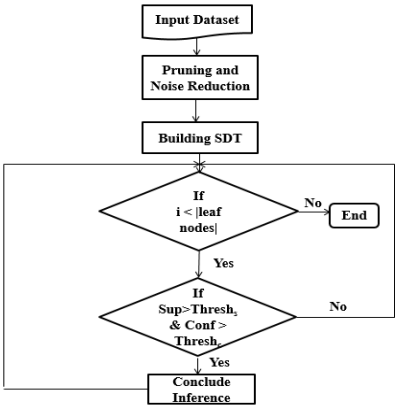
where $p_i$ is the number of tuples in the parent node in the decision tree, $N$ is the total number of tuples and $n_i$ is the number of tuples in the present node. Two thresholds are maintained, $Thresh_s$ and $Thresh_c$ for support and confidence respectively. For each node, if the *support* and *confidence* value exceeds the corresponding threshold, the inference is confirmed. The corresponding algorithm of evaluating the *support* and *confidence* for each candidate inference is described in Fig. 1(a). Fig. 1(b) describes the overall flow of the accident pattern inferring framework. It contains the following steps,

- *Handling the input dataset*
  The input dataset is a collection of tuples, describing the nature of an accident with its surrounding condition with some attributes. One of these attributes is chosen as the target attribute, whose behaviour is to be predicted.
- *Pruning and noise reduction*
  The input dataset may contain many incomplete, spurious information. These information will hamper the proper functioning of the algorithm and can lead to improper inferences. Thus, the data has to be pruned and trimmed to reduce noise. In order to prune the data, incomplete tuples are identified and removed. After pruning, few more refinement steps follow. In the given input data, each attribute can have multiple

**Algorithm**:Drawing Inferences from Decision Trees
**Input**: Classical and spatial decision trees
**Output**: Implicit accident patterns

```
foreach node do
    n_i=number of tuples in this node
    p_i=number of tuples in parent node
    Support = p_i / N
    Confidence = n_i / p_i
    if support > Thresh_s and confidence > Thresh_c then
        Inference Drawn
    end
    else
        Discard
    end
end
```

(a) Algorithm for evaluating *support* and *confidence*




(b) Flow diagram of accident pattern inferring framework

Fig. 1: Algorithm and flow diagram for inferring accident patterns

values, and the frequency of some attribute values can be very less. In order to overcome this problem, clustering has to be applied on the specialized categories of the attributes. Clustering involves grouping of two or more values of any attribute into one higher level value. This is how the input dataset can be standardized based on the required application.

- *Inference drawing*
  After pruning and refinement of the data, classical and spatial decision trees are built. On the basis of the support and confidence evaluations, inferences are drawn from the *decision trees*.

## IV. Experimentation and Result Analysis

The experimentation have been carried out on the accident data, collected on the National Highway(NH6), between Kharagpur and Kolkata, India. This route endures a lot of traffic and is also prone to over-speeding vehicles, thus increasing the chances of accidents. Modeling the pattern of accident data and their analysis is important because it leads to the surfacing of various conditions, leading to the fatal accidents. Another important application can be in identifying accident hotspots, or areas where grievous accidents occur frequently and to investigate the reasons behind such accidents. All these inferences can be extracted from the *spatial decision trees* as it encompasses the spatial aspect of the information. The dataset for this empirical experimentation, along with its schema, is specified in Table I.

### A. Inferences

The *spatial decision trees* are formulated from three years dataset. It gives a proper insight to the huge spatial dataset. Some useful inferences are extracted and discussed, which are implicit to the data. Each of the inferences is followed by a portion of the *decision tree* providing evidence for the same.

- **The rate of *Fatal* injuries increased in 2013 compared to 2012**
  From the dataset over the consecutive years, it was noticed that the number of *Fatal* accidents in 2013 was more than that of 2012.

- **All the *Fatal* accidents resulted from *Right_hand_side_collision*, *Rear_end/ Side_brush_collision***
  The nature of accident which resulted in *Fatal* consequences was *Right_hand_side_collision*, *Rear_end/ Side_brush_collision*. Fig. 2 and Fig. 3 show the portion of the *decision tree* corresponding to the above inference.

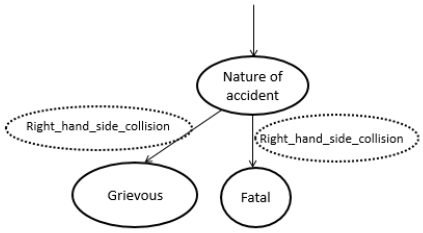



Fig. 2: *Fatal* accidents due to *Right_hand_side_collision*

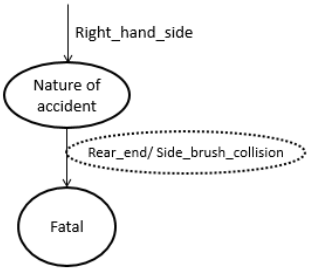



Fig. 3: *Fatal* accidents due to *Rear_end/ Side_brush_collision*

TABLE I: Training and test dataset attributes with possible values of the *spatial decision tress*

| **Supporting attributes** | **Possible values** |
|---|---|
| *Time of accident* | Day, Night |
| *Weather conditions* | Good: Normal;<br>Problematic: Very_hot, Rain;<br>Worse: Dense_fog, Dust_storm |
| *Side* | Left_hand_side, Right_hand_side |
| *Road condition* | Sharp_curve, Straight_road, Bumps, Dips |
| *Nature of accident* | Skidding, Rear_end/ Side_brush_collision, Right_hand_side_collision, Head_on_collision |
| **Class label attributes** | **Possible values** |
| *Type of accident* | Fatal, Grievous, Major, Minor, Non-injury |
| **Spatial attributes** | **Possible values** |
| *Location (distance of the accident from a reference point)* | $\mathbb{R}^+$ |

- **The effect of *Skidding* resulted no severe injuries like, *Fatal*, *Grievous*, *Major* injury**
  From the *decision trees* computed for the accident data of 2013, it was seen that *Skidding* resulted in less severe injuries, like, *Minor_injury, Non-injury* when compared to the other types of accidents. It is shown in Fig. 4.

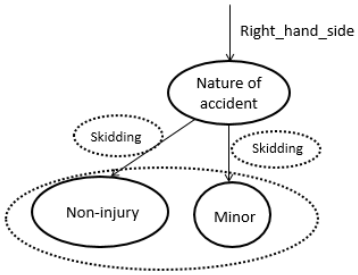



Fig. 4: *Skidding* resulted in less severe injury (*Minor_injury*, *Non-injury*)

- **In 2012, *Skidding* resulted in *Fatal*, but in 2013, no *Skidding* resulted in *Fatal***
  Fig. 5 and Fig. 6 show that in 2012, all *Skidding* accidents resulted in *Fatal* injuries, but in 2013, they were all non-*Fatal*.

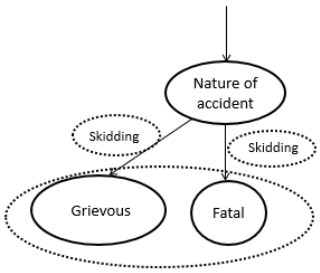



Fig. 5: *Fatal* injuries due to *Skidding* accidents in 2012

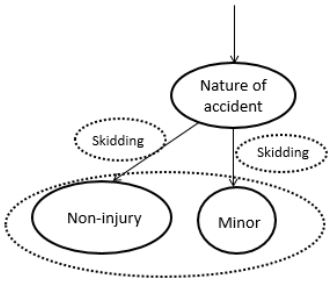



Fig. 6: *Non-Fatal* injuries due to *Skidding* accidents in 2013

- **All the *Fatal* accidents in 2012 happened during *Day* time**
  It is evident from the *decision tree* that in 2012, all *Fatal* accidents were accumulated during *Day* time.

## V. Conclusion

This paper aims at applying *spatial decision trees* to incur useful inferences from the real world accident data. This technique gives an insight into the patterns which are not explicitly evident from large spatial datasets. On the basis of the inferences drawn from these *decision trees*, accident management can be improved. Various conditions and patterns could be identified which lead to accidents in some locations. This information can be helpful to prevent accidents, take some corrective measures. Accident hotspots can also be identified by the *spatial decision trees*, and can be given prime importance to reduce the severity as well as the frequencies of the accidents in future. This can be useful in case of road traffic analysis and identifying implicit patterns in the traffic flow. Seasonal variations and its effects on the accidents can be considered as the future prospect of this work.

## References

[1] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees belmont. CA: Wadsworth International Group (1984)

[2] Quinlan, J.R.: Induction of decision trees. Machine learning **1**(1) (1986) 81–106

[3] Chang, L.Y., Chen, W.C.: Data mining of tree-based models to analyze freeway accident frequency. Journal of Safety Research **36**(4) (2005) 365–375

[4] Chang, L.Y., Wang, H.W.: Analysis of traffic injury severity: An application of non-parametric classification tree techniques. Accident Analysis & Prevention **38**(5) (2006) 1019–1027

[5] Ester, M., Kriegel, H.P., Sander, J.: Spatial data mining: A database approach. In: Advances in spatial databases, Springer (1997) 47–66

[6] Zhang, J., Guo, D., Wan, Q.: Geospatial data minig and knowledge discovery using decision tree algorithm -a case study of soil data set of the yellow river delta (June)

[7] Sitanggang, I., Yaakob, R., Mustapha, N., Nuruddin, A.: An extended id3 decision tree algorithm for spatial data. In: Spatial Data Mining and Geographical Knowledge Services (ICSDM), 2011 IEEE International Conference on. (2011) 48–53

[8] Chelghoum, N., Zeitouni, K., Boulmakoul, A.: A decision tree for multi-layered spatial data. In: Advances in Spatial Data Handling. Springer (2002) 1–10

[9] Ghimire, B., Bhattacharjee, S., Ghosh, S.K.: Analysis of spatial autocorrelation for traffic accident data based on spatial decision tree. In: Computing for Geospatial Research and Application (COM. Geo), 2013 Fourth International Conference on, IEEE (2013) 111–115

[10] Claramunt, Christophe: A spatial form of diversity. In Cohn, A., Mark, D., eds.: Spatial Information Theory. Volume 3693 of Lecture Notes in Computer Science. Springer Berlin/Heidelberg (2005) 218–231 10.1007/11556114_14.