A Novel Approach to Analyze and Predict Aircrash in Aviation

N. Sivaram  
M Phil Research Scholar, School of IT and Science  
Dr. G R D College of Science  
Coimbatore, India

T. Santha  
HOD, School of IT and Science  
Dr. G R D College of Science  
Coimbatore, India

Abstract—Aviation is one of the critical modes of our national transportation system. As such it is essential that new technologies be continually developed to ensure that a safe mode of transportation becomes even safer in the future. Data analyses in aviation improve safety. Analyzed air crash details play very important role to avoid further unforeseen accidents. Air crash details of data's are classified based on the parameters like longitude, latitude and country, which helps to analyze the flight departing. After analyzing the data of flight details will be analyzed based on clustered classes using cart algorithm. The data pertains to accidents involving flights within database were analyzed against accident databases and the results were compared. Decision tree drawn to get the analyzed data in increasing order ranked the findings by the factor support ratio, the result are displayed revise in text format.

Keywords—aircraft, accident analysis, database, crew reports, Pilot, air traffic control, aviation safety, knowledge discovery.

1. INTRODUCTION

An aircraft accident is associated with the operation of errors using an attribute focusing Technique to identify anomalies in distribution of different attributes of erroneous operations. Aviation data analysis is faced with in the issue of finding analysis methods that are effective at finding interesting patterns in the collected data that can be performed with limited computational resources. The analyzed data consists of reports on erroneous operations only. In addition such analysis must not require expertise in statistics. There are no similar data collected for the flights that completed without any operational errors [1]. Typically the process of analyzing the data is not very automated, analysis process often relies heavily on the past experiences. Airline accidents focusing on pilot errors, human cognition in responding to different situations and suggested accidents are caused by confluence of multiple factors, no systematic trends were found in the accident dataset while performing a trend analysis, research has analyzed larger sets of data available on accident to determine the casual factors of accident and incident.

Data mining tools do not replace the report of human analysis that helps the analyst with the analytical resource of task. Analyzing both accident and incident data to show the relationships between the events of two classes and to identify factors that are significantly associated with incidents. Aviation collects and analyzes the accident reports by a combination of automated and manual methods. Analysis of data is done by who are familiar with the domain.

Data Mining Workbench is a tool it is designed to analyze collection of available safety data. It has particularly complementary of other tools often employed, such as incident tracking systems or exchange data with others. An unintended consequence of disturbances to the flight schedules due to unpredictable circumstances such as bad weather or aircraft malfunctions, the regular flight schedules sometimes causing chaos at airports and airline operation centers [2]. The workbench is to find subtle patterns in the accident reports of safety risks before these risks lead to accidents. Tools like query capabilities and histograms can help to focus the intensive wide range of capabilities.

ACCIDENT TRACKING FLOW
II. WORK BENCH

In Data mining technique first is called Find Similar and is designed to search a collection of accidents to find those most similar to selected accidents. This is useful in determining if similar incidents or accidents have occurred before and if so how were they addressed. Find similar uses both structured and unstructured text fields to calculate the similarity between records.

I. FIND SIMILAR

Find similar search a collection of accident records to find those most similar to a selected target accident record. Technique was developed for find similar is a hybrid approach motivated by a need we found when discussing data mining with airlines safety. One task that was repeatedly called on to perform is to find records of accidents that are similar to those that just recently occurred. If the new event is found to be similar with some past records, it may be part of a larger and more serious pattern[2]. If on the other hand, the accident is anomalous it may be noted and closed or simply announced to the relevant department as a warning.

Determination of record similarity was not well supported by tools available to the safety. We could perform queries on both the structured and unstructured free-text fields, but these only respond with exact matches. Similarity of match between entire accident records was not supported. When doing this match, we use methods that are appropriate for the type of attribute.

\[
\text{Sim( recordi, recordj )} = w1*\text{match( x1i,x1j )} + w2*\text{match( x2i,x2j )} + \ldots + w_n*\text{match( xni,xnj)}
\]

\[
\text{Match (xni,xnj)} = 1 \text{ if x1i = x1j} \quad \text{Else} = 0
\]

When the data records are ordered the system requires information from the user concerning the size and also ordering of the domain. This matching is appropriate for any ordinal or interval type of data from numeric e.g. (Phase_of_flight) to string-based (Event Id, Error, Investigation Type, Location, Event Date, Country) given |Domain x| the match function.

\[
\text{Match (xni, xnj)} = 1 - (xni - xnj) / |\text{Domain x}|
\]

When the data are textual, vector space matching is used. In this approach a vector with length equal to the size of the vocabulary is built for each field. The value at position x represents the ratio of the number of times that word appears in the document term frequency and the number of times that word appears in the collection document frequency.

III. FIND ASSOCIATION

Find Association is a technique searches the collection of accidents to find subsets that have an interesting correlation. For example this tool can identify a set of accidents that share a common location of longitude and latitude and accident type. Knowing that such an association exists between accident types, locations, and aircrafts may help in determining what action to take to reduce or eliminate those accidents in the future.

Find Associations finds co-occurrences of different data values which finds associations among data values. For eg: one finding or association rule might be

Bird strikes -\rightarrow March 6\%: 72\%

Meaning there is an association between the bird strikes and the month of March. The values at the end of this association rule detail the strength of that association. The first value (6\%) describes the support of this rule. Support is calculated as the number of times the two or more values on the left hand side must co-occur before they are considered for a candidate association rule. The second value (72\%) describes the confidence of this rule [3]. The confidence is calculated as the ratio of how often the entire rule is true over the frequency of the left-hand side of the rule. In this example confidence is 72\%, meaning the ratio of bird strike accidents in March over the total number of Bird strike accidents is 72\%. The user provides minimum thresholds on both confidence and support to guide the efficient search for associations.

A global terrain database with coverage is resident with using the input latitude, longitude, altitude as well as flight path angle, turn rate and groundspeed, aircraft position within the terrain data and “look ahead” to potential conflicts with terrain.For the association A -\rightarrow B, Lift measure is calculated as \{((\text{Confidence of A -\rightarrow B}) / (\text{Frequency of B}))\}.

IV. FIND DISTRIBUTION

Find Distributions is a technique focuses on a selected field or attribute of the accidents. It first determines an overall distribution for these fields, subsets of the data are then obtained and the distribution of the selected field is calculated for each subset. Those subsets that differ most from the overall distribution are identified as the most interesting. This technique helps in identifying candidates for action.
It focuses on a user specified attribute to find exceptions to the common distribution. This tool uses an attribute focusing technique and works on the structured fields in the data only.

It compares the distribution with records the target and the match describe getting lost after running into bad weather. There is an important difference between the two matches found. The match on the text and structured fields. (A) Found a record of an accident that occurred in the daylight while the match on text only fields. (B) Found an accident that occurred at night. This example shows how information from both text and fixed fields can be exploited to provide a better overall match between accident reports than using just text or just fixed fields alone these techniques are automated to do certain analyses that are either difficult or resource intensive.

V. CONSTRAIN DATA

Accident data that can be assumed complete and free of bias. Data that however is under-reported and subject to self-reporting bias. To address these constraints underlying factors of accidents and incidents qualitatively. The historical data on accidents is large enough to represent these factors quantitatively. Also, we consider all factors that have been present in an event, regardless of their primary or contributory role in leading to the event. This minimizes the impact of the bias in reporting the factors of constrains.

VI. DATA PREPARATION

Normalized the data across the databases and then developed ontology by developing a hierarchy of factors and sub factors common across the databases. Normalization of the values was needed so that all databases use the same term to refer to the same factor or condition [5]. Transformed the reports into vectors consisting of fields that indicate absence or presence of each of the common factors and sub-factors in the event accident or incident. Then analyzed these vectors.

i) Airport - Snow not removed from the runway, poor Lighting, confusing marking.
ii) Air traffic control - Communication with pilot, Complying with procedures.
iii) Aircraft - Flight control system, Engine, Landing Gear.
iv) Maintenance - Compliance, Inspection.
v) Weather - Wind, Thunderstorm, Ice.
vii) Other - FAA oversight, Visibility.

VII. DATA ANALYSIS

Applying the CART algorithm to perform sets of data analyses. In each analysis the incident vectors were paired with accident vectors from one of the databases. Each analysis identified patterns of factors which are significantly associated with accidents.

We ranked the findings of each analysis using the factor Support Ratio measure. Final results of the analyses were compared at the end.

VIII. ALGORITHM

CART is classification method which uses historical data to construct decision trees. Depending on available information it finds conjunctions of attribute value pairs that are significantly different across multiple nodes [6]. In the case of data, there are classes: incident and accident, longitude and latitude vectors. Attribute values are binary values (0 or 1) for the factors in each event vector, implying presence or absence of the factors in that event.

Input accident vectors
C = set of deviations, initially empty
D = set of factors in the input vectors
Scan input data and count support
suppacc= (incidents containing the factor/total incidents)*100
suppac= (accidents containing the factor/total accidents)*100
library ( " xclust " )
2 boston = read (" bostonh ")
3 data = boston [1:100,]
4 Var = data [,1:13]
5 Class = data [,14]
6 x = data [,1:2];
7 y = data [,3]
8 NumberOfLines = rows (data )
9 tr = cartsplitregr (Var , Class , 30)
10NumLeaf = cartleafnum (tr )
11 NumLeafproc (y ) = simulate (seed, n )
12 randomize (seed)
13 cartdissptree (tr )
// generating data with layout
xdat = uniform (n ,2)
index = ( xdat [,2]<=0.5) + ( xdat [,2]>0.5) *
(xdat[,1]<=0.5) + 2.*( index ==0) + 3.*( index =2)
// return the list
16 x = data . xdat
17 y = list ( xdat , index , color , layout )
18 Generate children (factor-set, C)
19 For each child
   tr = cartsplitclass ( newx , newy , 0, 1);
20 predictedClass = cartpredict ( tr , testingData )
21 if ( predictedClass == realClass )
22 CorrectAnswers = CorrectAnswers + 1
23 CorrectAnswers / NumberOfLines
24 endp

Legend:-

Suppinc = incident support
Suppacc = accident support
dev = deviation

Factor sets with a deviation is relaxed for child generation.
The reason for this modification is to allow for discovery of factors that might not be individually associated with incidents but if combined together they could be significant factors. Discovery of such cases is one of the objectives of the analysis.

IX. RANKING

Factor sets that are significantly identified by the algorithm, we rank them by their Factor Support Ratio measure. Calculate the Factor Support Ratio for each factor set as the probability of the factor-set given an accident, divided by the probability of that factor set given an accident, or the ratio of the factor set’s support in accident dataset over its support in the incident dataset where $F =$ factor set, $acc =$ accident, $inc=incident$, $P(F|acc) = probability$ of factor-set given an accident, $#Facc =$ number of accidents containing factor $F$, $acc =$ total number of accidents.

\[
\text{Support Ratio} = \frac{P(F|inc)}{P(F|acc)}
\]

\[
= \frac{P(inc/F) P(F)/P(inc)}{P(acc/F) P(F)/P(acc)}
\]

This information conveyed by the Support Ratio measure about the factor set is different than the deviation that is used in the algorithm [7]. Deviation is the differ between the factor set’s incident and accident supports. Support Ratio is the probability of a factor-set being involved in an incident divided by its probability of being involved in an accident.

X. RESULTS

Results of the analysis were consistent with findings and some were interesting in previous studies had not identified them.

Highest ranked accident factors were Air Traffic Control (ATC) followed by the pilot factors. Among the ATC factors, communications sub-factor had the highest rank of association with accidents [8]. And among the pilot factors, visual look out had the highest rank. Identification of ATC communications and pilot visual look out as accident factors was consistent with previous findings. The interesting finding was that ATC factors which are less frequent than pilot factors were ranked higher. This implies that although ATC factors are less frequent but once they occur there is a high risk of having an incident as opposed to an accident. Pilot factors are more frequent than other factors in accidents but they are also more frequent in incidents, which makes their Support Ratio lower and ranks them after the airlines.

Association of another finding is aircraft factors with incidents. Aircraft factors are mechanical problems with the aircraft system or components, such as landing gears and flight control systems. The results showed these factors are more likely to be involved in incidents except when combined with other factors such as pilot errors or severe weather [9]. Airline factors such as mistakes by the airline or company personnel and inadequate or lack of procedures by the company for performing a task were consistently the highest ranked category of factors associated with accidents among the high-level categories of factors its an interesting result.

Research factors across multiple databases and in addition to identifying the factors associated with the accidents we could rank in order of their significance. Results are associations that were consistently identified by multiple analyses. Additional associations were identified by each individual analysis [10].

<table>
<thead>
<tr>
<th>Database</th>
<th>Support ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>weather, pilot</td>
<td>1.9</td>
</tr>
<tr>
<td>airport, pilot, other</td>
<td>2.3</td>
</tr>
<tr>
<td>aircraft, pilot, weather</td>
<td>2.9</td>
</tr>
<tr>
<td>company, pilot, other</td>
<td>3.6</td>
</tr>
<tr>
<td>aircraft, pilot,company,other</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Aircraft - (aircraft) -> Incident
(Aircraft, weather) -> Accident

ATC - (ATC) -> Accident
(ATC, airport) -> Accident
(ATC, pilot, airport, other) -> Accident

Pilot - (pilot) -> Accident
(pilot, weather) -> Accident
(pilot, airport, other) -> Accident

Ranking of the results when multiple factors are present, there is a higher likelihood of having an incident, signifying that pilot factors combined with airport factors are more likely to result in accidents than the pilot factors alone.
XI. CONCLUSION

Applying CART algorithm to the aviation safety data set, we can able to analyze aircraft accident data in contrast to the longitude and latitude data and identify patterns of factors which are associated with the accidents. Our ranking measure, the Factor Ratio, allowed ranking of the findings and identification of relative significance of the factors in contributing to accidents, compared to other factors. This work analyzed aircraft accidents and pertaining to commercial flights. The methodology used here could be applied to the general airlines as well. The analysis could be extended to include worldwide safety events.

REFERENCES


AUTHORS PROFILE

N. Sivaram received his B.sc in Department of Mathematics from Bharathiar University, Coimbatore and Postgraduate (MCA) in Department of Computer Application from Bharathiar University, Coimbatore. He is doing M.Phil, Research Scholar of IT & Science in Dr. G R Damodaran College of Science, Coimbatore. His research interests include Data Mining algorithms and its Complexity issues.

Dr. T. Santha, HOD, School of IT & Science, Dr. G R Damodaran College of Science, Coimbatore, M.sc(Maths), M.phil(Maths), M.phil(CS), Ph.d.