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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

Dempster Shafer Theory Based Uncertainty Quantification in Air Traffic Management

Deepali Pande[1], M. Sondawale[2], Gunjan Keswani[3], P.Walke[4],

Assistant Professor, Department of Computer Application, RCOEM, MH.India [1, 2, 3]

Assistant Professor, Department of Computer Application[4], SVPCET, MH.India [4]

ABSTRACT: The intention of this paper is to examine data attained from the U.S. Department of Transportation, Bureau of Transportation Statistics on airline flights. The data is based on domestic passenger flights within the United States of America between US airports. A system has been generated around this data to allow insights and gain knowledge from the downloaded data. We use Dempster Shafer evidence theory in calibration of flight plan delay over raw dataset acquired from heterogeneous sources of from infrared sensors. The types of insights that have been addressed in this system are airports and airlines performance and league tables of airlines to see how they perform from month to month. The results allow the user of this application and reports to see if certain airlines or airports perform better than others and perhaps allow the user to choose wisely between alternatives based in the knowledge gained.

KEYWORDS: Dempster Shafer Theory, evidence, flight plans

I. INTRODUCTION

Flight delays are a simple fact of life and affect most people at some time or another. Moreover, if one has to link to a connecting flight or if an important meeting is missed then the issue is compounded. If the delay is small enough it can pass un-noticed but if it is significant it can have a cost attached to it. Personal costs would fall under accommodation costs or finding alternative travel arrangements, as for airlines it would include crew costs and passenger costs in having to re-accommodating them along with airport costs and rescheduling costs of flights. It is estimated that delays cost the US economy in the region of 16 Billion dollars a year due to the delays experience by airline and airport users according to studies carried out by NEXTOR, 2010.

The premise of this project is to build an application that can generate number of reports pertaining to flight delays. It is based on data around US domestic flights. The data has been attained from Bureau of Transportation Statistics. It is intended to show the results to the user using graphical methods including some tables and charts.

Using this application, number of reports will be compiled to show how the airlines and airports perform. It will show the user how over time changes occur and display the results via table and charts.

II. RELATED WORK

The data has been taken from the Bureau of Transportation Statistics site. This site holds data which originates from the Research and Innovative Technology Administration, (RITA, 2015) site which among its numerous data sets hosts this particular data set.

The data is based on a collection of domestic flights within the United States of America. It holds these details for both departures and arrivals and gives further data regarding airlines, airports and other metrics recorded against these flights.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

In all it has 29 columns of data regarding the various aspects of each flight that has flown over the duration being examined.

A lot many number of years of data held on this site from 1987 to 2015. Each year's data set was quite large, being in the region of 70 to 80 MB and containing approximately 5 to 6 million entries. One entry per flight over the period of time has being examined.

Other data is used providing translations from airline and airport codes to their respective names. Other details regarding the state and city, the airport served is held in this supplementary data. These were also downloaded. Using these we could provide further analysis capability by attaching airports to the respective states and seeing if grouped certain states have better performance than others.

III. OBJECTIVE

Delays in flights are a common occurrence but what is the typical delay and how are these dispersed between airlines and airports? Are there specific States that experience these, or are they spread evenly over the whole of the country? What are the extremes of the delays incurred? Can you avoid them by choosing wisely between certain airlines and destinations that do not perform particularly well? These are the typical questions that travellers could well ask themselves.

Building a system that will take the data and incorporate it into a number of reports will show if indeed patterns are evident from within the data. The results returned will empower the user with knowledge on flight delays and knowledge is power. Through the knowledge gained from the resulting reports, it could empower the user to make a more informed judgment on the route or airline chosen to travel. Along with the travellers, the airports and airlines could also benefit from the knowledge. Showing them if improvements can be made and identifying where they need to make these improvements.

The typical traveller using flights is aware of delays and the costs associated and the degree to which they should be compensated in the event of such delays. Airlines themselves are only too aware of the associated costs that arise when the delays are incurred. Thus passengers can choose more optimal routes where performance is satisfactory and conversely airports and airlines can see focus areas where they need to make improvements.

IV. EXISTING SYSTEM

There is currently considerable enthusiasm around the MapReduce (MR) paradigm for large-scale data analysis. Although the basic control flow of this framework has existed in parallel SQL database management systems (DBMS) for over 20 years, some have called MR a dramatically new computing model. Like MapReduce, they provide a high-level programming environment and parallelize readily. Though it may seem that MR and parallel databases target different audiences; it is in fact possible to write almost any parallel processing task as either a set of database queries or a set of MapReduce jobs.

V. MATHEMATICAL FORMALISM OF DEMPSTER SHAFER THEORY

The Frame of Discernment: A set which is represented by θ contains all possible elements in every context and also all the possible elements of interest are exhaustive and mutually exclusive events. Such a universal set is called the frame of discernment. In theory of probability, it is simply known as sample space. In the example of tossing a coin, the frame of discernment will be {T, H} where 'T' represents 'tail' and 'H' represents head. Here, in Dempster Shafer theory of evidence, the frame of discernment is the set of all hypotheses taken into consideration.

The Basic Probability Assignment function (BPA): If we have θ as the frame of discernment then the function $m = 2^{\theta} \rightarrow [0, 1]$ is called the basic probability assignment if m (Φ) = 0 and

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

$$\sum_{A\subset\theta} m(A) = 1$$

Where the term m (A) is the measure of belief committed to 'A' exactly and the term m (A) is called the basic probability number of 'A'.

The Belief Function: The belief function Bel (A) measure the total belief of all possible subsets of A. It is also called m function. It is represented on $m = 2^{\theta} \rightarrow [0, 1]$:

$$\text{Bel(A)} = \sum_{B\subset A} m(B)$$

The Plausibility Function: The Plausibility function is thus represented from belief as

$$\text{Pl(A)} = \sum_{B\cap A\neq 0} m(B)$$

Here, Pl (A) = 1 – Bel (A) and Bel (A) ≤ Pl (A)

The Doubt Function: It is simply defined over the complement set A as
Doubt (A) = Bel (*A*).

The Rule of Combination of Evidences: The Dempster Shafer's Rule of combination is just like a fusion operator which derives commonly shared beliefs from heterogeneous sources. In this process, all unshared beliefs are removed through a normalization factor. The combination also called the 'joint mass' is calculated from the two sets of masses namely m1 & m2 as follows:

$$m_{1,2}(\emptyset) = 0$$

$$m_{1,2}(A) = (m_1 \oplus m_2)(A) = \frac{1}{1-K} \sum_{B\cap C = A\neq\emptyset} m_1(B)m_2(C)$$

Where 'K' represents the quantity of conflict between the two sets of masses.

$$K = \sum_{B\cap C=\emptyset} m_1(B)m_2(C)$$

VI. UNCERTAINTY QUANTIFICATION USING DEMPSTER SHAFER EVIDENCE THEORY

The Dempster Shafer theory as developed by Arthur P. Dempster & Glenn Shafer provides a general framework to model hypothesis in concept of epistemic uncertainty. The theory helps in combining evidences from different sources to transform it into a mathematical object called the belief function. In practice, many authors have proposed theories to quantify uncertainty but DS theory provides various properties and functions to range the hypothesis in terms of belief and plausibility intervals.

In Dempster Shafer Theory, the belief function is used to measure the confidence of the occurrence of an event in the given hypothesis. Belief functions base upon the degrees of belief (or confidence) on a set of related hypothesis. In formalism, the DS theory has vivid applications in sensor data fusion. Sensory data from heterogeneous sources can be combined to formulate a good measure of probability. The data chunks obtained from different sources of sensors individually have strong uncertainty ratio when it is available in a raw form. These data chunks are fused to obtained big pieces of data set using sensor fusion concept as supported by Dempster- Shafer theory. The data set derived in this manner thus has reduced ratio of uncertainty and is useful for quantification. The data sources for a fusion process have been originated from heterogeneous sources. The method of indirect fusion is used to relate the data set from a priori information acquired from the heterogeneous sources. The degrees of belief are thus based upon independent items of evidence in each chunk of the dataset. In implementation, the degree of belief which is also depicted as a mass is pictured as belief function. Probability values are assigned to each set of possible outcomes in the Arrival and departure flight plans in the data set.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

According to the formalisms described in the Dempster- Shafer theory, assumed masses are assigned to all the non-empty subsets of the propositions which comprise of the entire data in the hypothesis. Using the framework of Shafer's theory, belief about the proposition are represented as intervals with two bounding values belief (or support) and plausibility where **Belief ≤ Plausibility**

EXAMPLE OF FLIGHT DELAY CALCULATION USING DST

The raw chunks acquired from the flight dataset are used for formulation of belief and plausibility functions as described in the section III. Suppose a belief of 0.2 and plausibility of 0.7 is evaluated for the flight delay determination proportion with hypothesis saying "The Flight Plans are late today". Here it can be justified to have evidence strongly stating with a confidence of 0.2 and a true value. But, the evidence contrary to the hypothesis that "The flight plans will not be late" has a confidence only of 0.3 according to the relation that plausibility is one minus the sum of masses of all sets whose intersection with the hypothesis is empty. The interval of determination in the hypothesis is 0.1 which states the difference between the belief and plausibility. This interval thus represents the level of uncertainty determined based on the evidences in the system.

VII. ESTIMATING THE BELIEF FUNCTION FOR THE FLIGHT DATASET

1. The Sensor Mass Function is used to obtain context value of belief. The sensor mass functions assign mass to the context values from the 3D airplane trajectory in the following manner:
   [Actual Departure Time: 1; Actual Arrival Time: 0]
   [Scheduled Departure Time: 1; Scheduled Arrival Time: 0], where we use time as evidence factor.
2. Then the context value of belief is disseminated up to applicable situations using evidence propagation and complex relations as per the discussion in section V. The sensor mass functions assign mass to the context values from the 3D airplane and these values are propagated to form an extensive database of each time deviation ∂T(∂TD, ∂TA & ∂TETA)
3. A total belief function is computed using the theory of combining evidence for situations in the set using Dempster Shafer's evidence combination rule. After this, total belief is calculated for each situation using Dempster's Rule of combination. Here, we combine the averaged situation belief n-1 times where n is the total number of evidence sources for the situation.
4. Finally, the situation with highest belief is selected before ensuring non co-occurrence of situations.

VIII. PROBLEM FORMULATION

In application of Dempster Shafer evidence theory, we use arrival and departure times as the needful quantities. The actual parameters taken into consideration are actual departure time, actual arrival time, scheduled arrival time, scheduled departure time. The deviation of actual departure time from scheduled departure time is calculated as an estimate of interval. Hence, two quantities namely estimated time of departure and scheduled time of departure are available. Similarly, deviation of actual arrival time from scheduled arrival time serves as testament in calculation of interval of estimate for another two quantities namely estimated time of arrival and scheduled time of arrival.

We interpret these quantities as ΔTD and ΔTA for deviation in departure and arrival times. A posterior and prior dataset of multiple flight plans is used to describe uncertainty amongst the flight dataset used. The belief function for the interpreted quantities is formulated to describe uncertainty. The characteristics of belief functions are analysed using table 1 dataset. Figure 1 represents the system model which represents the method used in implementation. The Table 1 represents sample of calibrated data obtained from the approach as described in section VI.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

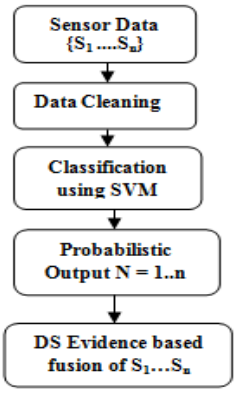

Figure 1. System Model

Flight	Phase	DT: Departure Time	Range of Uncertainty	Interval Size
Flight A	Up	17:10	0.25	0.22
	Down	13:20	0.11	0.09
Flight B	Up	18:15	0.16	0.14
	Down	20:10	0.06	0.04

Table 1. Uncertainty Characteristics of two flights

The software is developed using hadoop version 2.0+ with support of eclipse IDE on Red Hat Linux platform. The libraries used include JUNG Framework & JavaCSV. The performance is evaluated on dual core processor with 4 GB available RAM space. The language used for coding the system is Java as a Hadoop framework itself compiled in Java. The datasets used for comparison and analysis have been used from the Bureau of Transportation Statistics (BTS) website. The name of the dataset is 'On-Time Performance' freely available at http://www.transtats.bts.gov/ The 'On-Time-Performance' dataset has 109 different fields. For the purpose of this project 14 of these fields are retained from original dataset. The list of the required dataset fields is shown in
table 2.
The system ensures that the data that has been downloaded is clean. This involves inspecting the data in each file, removing any unnecessary fields and ensuring that it is compatible to be uploaded into the hdfs environment.

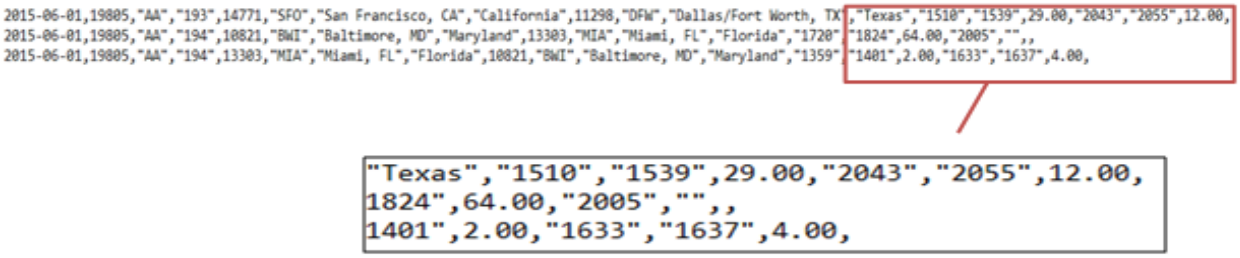

Figure 3. Uncleaned Data

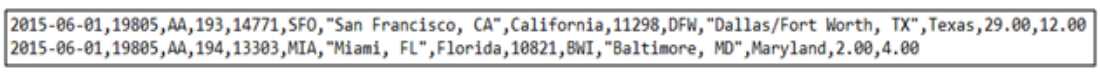

Figure 4. Cleaned Data

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

Sr. No	Fields	Description
1	FL_DATE	Flight Date (yyyymmdd)
2	AIRLINE_ID	An identification number assigned by US DOT to identify a unique airline
3	CARRIER	Code assigned by IATA and commonly used to identify a carrier.
4	FL_NUM	Flight Number
5	ORIGIN_AIRPORT_ID	Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.
6	ORIGIN	Origin Airport
7	ORIGIN_CITY_NAME	Origin Airport, City Name
8	ORIGIN_STATE_NM	Origin Airport, State Name
9	DEST_AIRPORT_ID	Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport.
10	DEST	Destination Airport
11	DEST_CITY_NAME	Destination Airport, City Name
12	DEST_STATE_NM	Destination Airport, State Name
13	DEP_DELAY_NEW	Difference in minutes between scheduled and actual departure time.
14	ARR_DELAY_NEW	Difference in minutes between scheduled and actual arrival time.

Table 2. Required dataset and its description

As we can see in figure 3 there is no entry for delay in the given dataset. So such records are removed from final cleaned CSV file, as we see in figure 4.

IX. RESULTS

Once the cleaned data is uploaded into hdfs, now the analysis can be done. As our system provides to gain insights of Airports as well as of Airline companies option should be provided to user to choose between them and also options of type of delay. On selecting the option, the appropriate analysis can be done. Figure 5 below shows this implementation.

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

Figure 5. Options for user

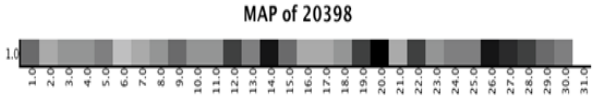

Figure 6. Heat Map

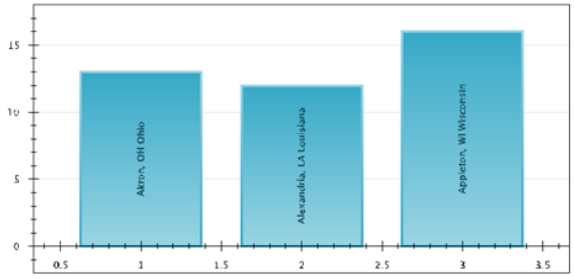

Figure7. Bar Graph

X. CONCLUSION

The procedure of Dempster Shafer theory of evidence was implemented for test on sensor data obtained from heterogeneous sources observed from infrared sensors. The procedure as described in the section VII was used to construct the belief functions. It is experimentally evaluated that the belief functions computed in the manner proposed show full accuracy. The two sets of flight plan data represent a cross latitude flight as obtained from the U.S. Department of Transportation, Bureau of Transportation Statistics on airline flights. A computational algorithm to implement the procedure was developed. This analysis has provided insights as to which carrier is the best to fly with, which carriers have the most delays, what airports have the most delays. These insights can only be used as a guide; one cannot do anything about severe weather disruptions or airport strikes. This analysis could also be of interest to

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

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 12, December 2015

airline carriers and airports as it will allow them to evaluate how they are doing in the marketplace. A carrier with the same amount of flights as another but experiencing much higher delays may be able to try and analyze what they are doing differently.

REFERENCES

[1] A. P. Dempster," Upper and Lower Probabilities Induced by a Multi valued Mapping", Annals of Mathematical Statistics, 1967
A. P. Dempster, "New Approaches for Reasoning Towards Posterior Distributions Based on Sample Data", Annals of Mathematical Statistics, 1996.
[2] G. Shafer, A Mathematical Theory of Evidence; Princeton University Press, Princeton, 1976
[3] F. Bosse, J. Roy, "Fusion of Identity Declarations from Dissimilar Sources Using the Dempster Shafer Theory", Optical Engineering, Vol. 36, No. 3, pp 648-657; March 1997.
A. Martin, A. Jousselme, and C. Osswald., "Conflict measure for the discounting operation on belief functions", In Proceedings of the Eleventh International Conference on Information Fusion, pages 1003–1010, Piscataway, NJ, 2008.
[4] L. Mathelin and M. Y. Hussaini, " A stochastic collocation algorithm for uncertainty analysis.Technical report," NASA, 2003.
[5] E. McConkey and E. Bolz. Analysis of the vertical accuracy of the ctas trajectory prediction process Technical report, NASA AMES, 2002.
[6] W. M. Moon, "Integration of geophysical and geological data using evidential belief function",IEEE Transactions on Geoscience and Remote Sensing, 18(4):711–720, 1990.