Fusion of Soft and Hard Data for Event Prediction and State Estimation

FUSION OF SOFT AND HARD DATA FOR EVENT PREDICTION AND STATE ESTIMATION

BY

ABIRAMI THIRUMALAISAMY, B.Eng.

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRICAL & COMPUTER ENGINEERING

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

Master of Applied Science (2015) McMaster University

(Electrical & Computer Engineering) Hamilton, Ontario, Canada

TITLE:	Fusion of Soft and Hard Data for Event Prediction and State Estimation
AUTHOR:	Abirami Thirumalaisamy B.Eng., (Electronics & communication Engineering) Karpagam University, Coimbatore, India
SUPERVISOR:	Dr. T. Kirubarajan
NUMBER OF PAGES:	ix, 62

To my Amma, Appa and Anna

Abstract

Social networking sites such as Twitter, Facebook and Flickr play an important role in disseminating breaking news about natural disasters, terrorist attacks and other events. They serve as sources of first-hand information to deliver instantaneous news to the masses, since millions of users visit these sites to post and read news items regularly. Hence, by exploring efficient mathematical techniques like Dempster–Shafer theory and Modified Dempster's rule of combination, we can process large amounts of data from these sites to extract useful information in a timely manner. In surveillance related applications, the objective of processing voluminous social network data is to predict events like revolutions and terrorist attacks before they unfold. By fusing the soft and often unreliable data from these sites with hard and more reliable data from sensors like radar and the Automatic Identification System (AIS), we can improve our event prediction capability. In this paper, we present a class of algorithms to fuse hard sensor data with soft social network data (tweets) in an effective manner. Preliminary results using are also presented.

Index terms: Dempster–Shafer belief theory, Random finite set theory, Modified Dempster's rule of combination, soft and hard data fusion, airborne surveillance of surface targets, event prediction, social data analysis

Acknowledgements

The successful completion of my research presented in this thesis could not have been accomplished without the help and support of a number of people. I am glad to have the opportunity to acknowledge them here.

First and foremost, I would like to express my deepest gratitude to my academic supervisor, Dr. T. Kirubarajan, whose expert advice and guidance motivated me to achieve excellence in my work. I would like to thank him for his valuable inputs throughout the course of my research.

I am most grateful to Dr. Tharmarasa and Ehsan Taghavi, for their good advice and support. Their thorough reviews and perceptive comments greatly improved the quality of my thesis work. I would like to thank Dr. Anne-Claire Boury-Brisset from Defence Research Development of Canada (DRDC) for her support throughout my research. I would also like to thank Dr. Sorina Dumitrescu and Dr. Jian-Kang Zhang for being members of my thesis defense committee. I appreciate all the time the members of my committee took to read this thesis and for providing their input and thoughts on the subject.

I am thankful to the Department of Electrical & Computer Engineering, for providing me an opportunity to pursue my research at McMaster University. I would also like to thank Cheryl Gies of the Department of Electrical & Computer Engineering

for providing me with prompt administrative support.

Last, but by no means least, I would like to thank my parents, my brother and my friends for their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice.

Contents

List of Figures

Chapter 1

Introduction and Problem Statement

In order to track and predict events, and to track mobile target states, military or homeland security systems need accurate data. Due to limited fields-of-view and obscuration, conventional prediction and tracking methods [3] that rely exclusively on hard sensors (e.g., radar, sonar, video) can make erroneous decisions. On the other hand, algorithms that use only soft data (e.g., human input, social network data) can be ineffective due to conflicting and unreliable information. In some cases, the unreliability of soft data might be intentional. Social Network (SN) data is one form of soft data that has many advantages: it is voluntary, voluminous, instantaneous and evolving. As a result, it is a rich source of data accumulated over time by a large number of identifiable users, who are often close to unfolding events of interest, at virtually no cost data. This has spurred great interest in mining social data for information extraction and exploitation. Specifically, the fusion of soft and hard data is of significant interest in many surveillance systems. This indeed provides the

motivation for the proposed hard and soft fusion technique.

1.1 Automatic Identification System

Automatic Identification System (AIS) is an autonomous and continuous broadcast system that exchanges maritime safety and security information between participating vessels and shore stations. AIS operates in the VHF maritime mobile band using Time Division Multiple Access (TDMA) technology to be able to meet high broadcast rates, while ensuring reliable and robust operation. Automatic Identification System provides a way for ships to electronically send and receive data, which includes vessel identification, position, speed and course with vessel traffic service stations as well as with other ships. AIS uses the Global Positioning System (GPS) [30] data over digital Very High Frequency (VHF) radio communication equipment to electronically exchange location as well as other information. AIS is generally used by marine vessels along with the Vessel Traffic Service (VTS) to monitor vessel location and movement, which is primarily needed for vessel traffic control, collision avoidance and other safety applications. AIS has previously been used in many applications including fusion with radar data, anomaly detection (see Figure 1.1) and traffic pattern analysis [36]. AIS used as the hard data source because of the focus on airborne surveillance of surface targets as well as its ubiquity and versatility. A typical AIS functionality as shown in Figure 1.2 uses an array of data collection aircraft and satellites as well as coastal stations to collect information about the movement of vessels. These data are collected and reached a central agency of the country from which it is shared between all other countries in the world.

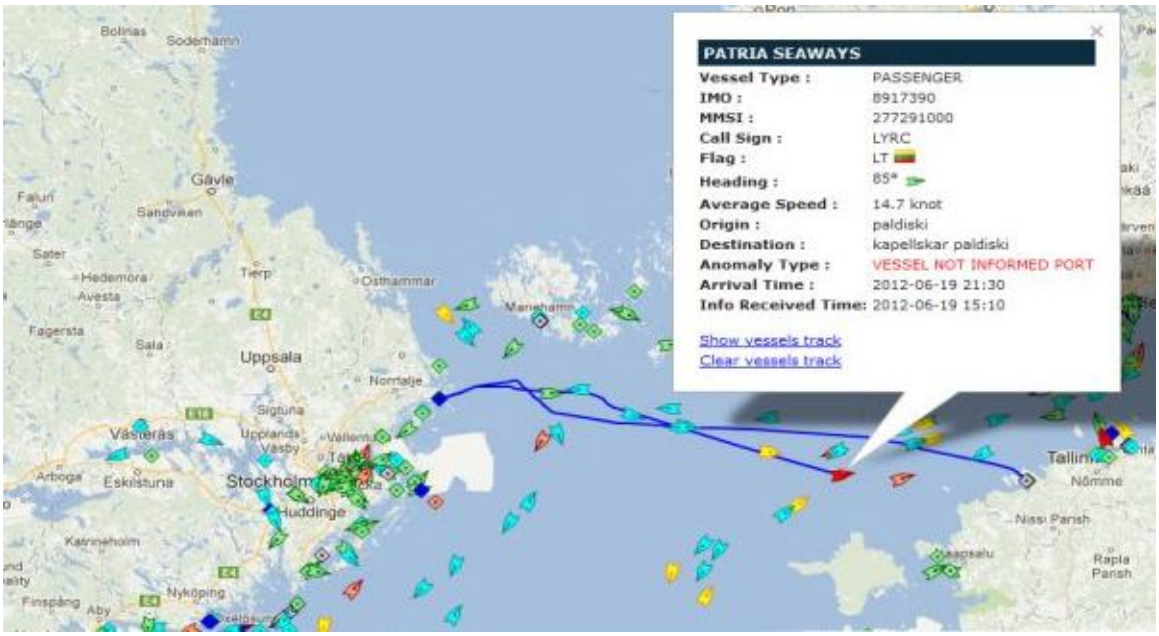

Figure 1.1: Anomaly detection based on automatic identification system data (courtesy of Google)

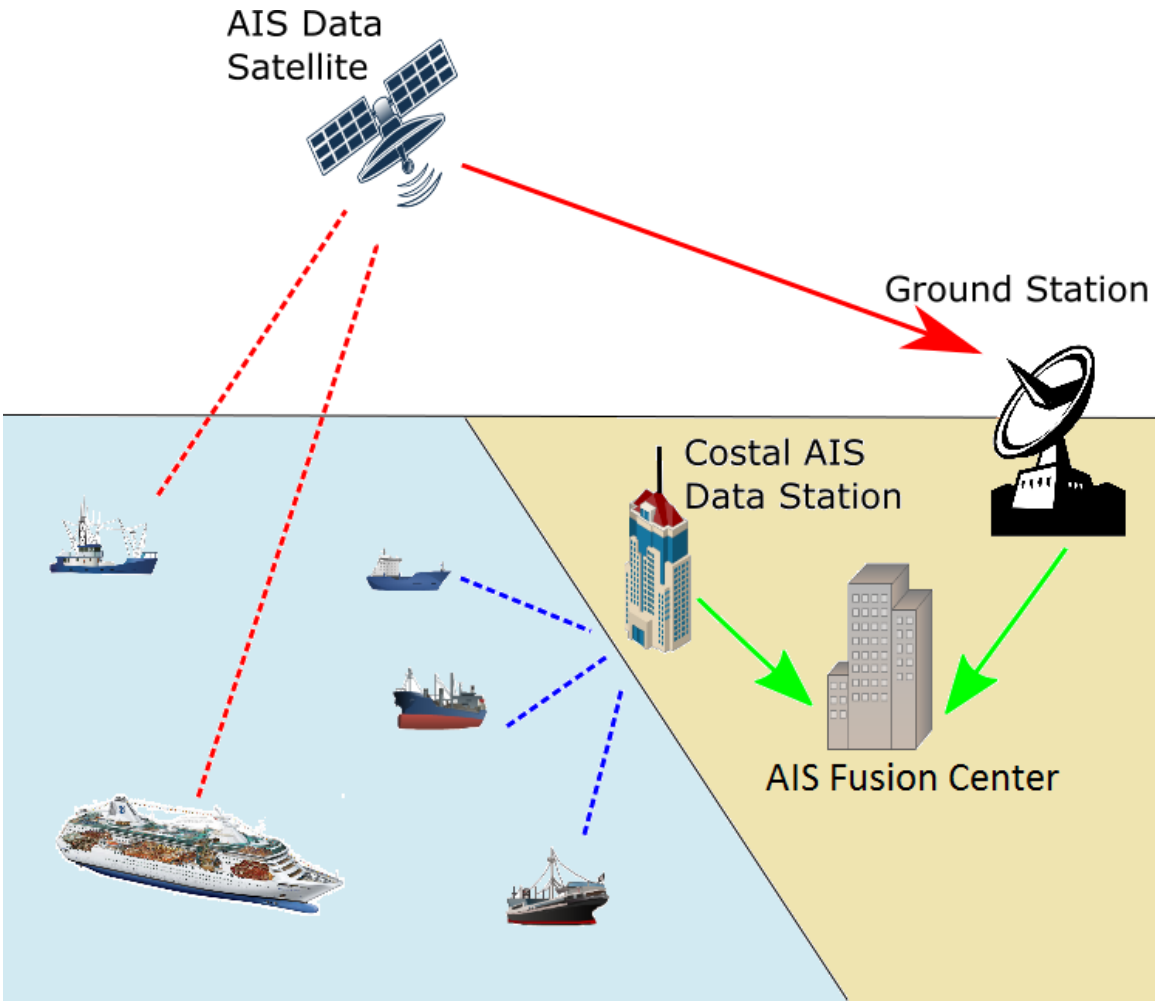

Figure 1.2: Automatic identification system functionality

1.2 Proposed Approach

The proposed approach follows a hierarchy in which the data from social networking sites such as Twitter is initially processed with a set of keywords as shown in Figure 1.3. The output of such a processing block is the refined information according to the list of keywords. The pre–filtered data is then sent to another block to be fused with AIS data. As a fundamental step, it is necessary to address the uncertainty in the soft data in order to convert it into a quantitative value. This step is crucial as the soft data is going to be compared against and fused with the hard data. In order to process large amounts of soft/hard data, an efficient method must be utilized. An efficient method to fuse the social network data with AIS data/hard data is proposed. The method followed in this thesis starts with processing the SN data with a set of keywords assuming that the keywords are defined based on some evidence (see Figure 1.3). After extraction, there will be uncertainties and conflicts in the collected data. To remove the conflicts and uncertainties we apply Dempster-Shafer Belief Mass Assignment to the data. Then, Modified Dempster's Rule of Combination (MDRC) [16, 11] is used to address the issue of fusing large amount of SN/Soft data with AIS/Hard data.

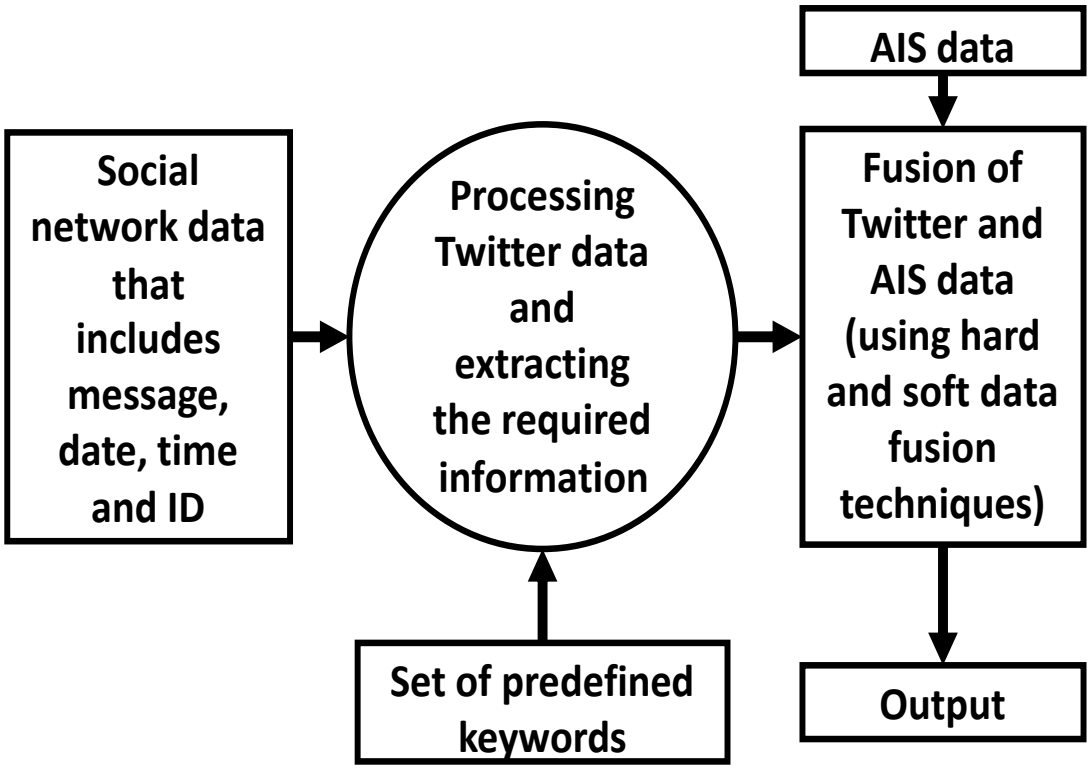

Figure 1.3: Block diagram of data flow in soft/hard data fusion

1.3 Motivation and Contribution

There are many challenges in the fusion of soft data with hard data since they are often incompatible with each other and the computational load of processing large amounts of social network data for fusion can be prohibitive. The incompatibility stems from the fact that soft data is qualitative while hard data is quantitative [33]. Both qualitative and quantitative information are needed to predict events or to estimate target states precisely and with real-time capability. Such fusion is of interest in asymmetric military operations where human–generated data are shown to be of crucial importance [1]. Recent developments in the literature on human–centered information fusion [51] as well as several preliminary works on soft/hard fusion are part of a trend towards more general data fusion frameworks [14], where both human (soft) and non–human (hard) data can be processed efficiently to yield better results. To develop an effective soft and hard data fusion system, one has to deploy an effective mathematical framework to fuse data and infer information while appropriately factoring in uncertainties. Commonly used frameworks for fusion are Dempster–Shafer, fuzzy set [52], possibilistic [7] and rough set theory [22]. This thesis presents a novel approach for fusing soft social network data with hard data. Specifically, Twitter feeds are used as the source of soft data while Automatic Identification System (AIS) [17, 40] reports are used as hard data. The context of the motivating problem is the prediction of events such as revolutions and terrorist attacks using social network data along with airborne surveillance data. The proposed work relies on the Modified Dempster's Rule of Combination (MDRC) [28] because of its simplicity and its ability to resolve conflicts during fusion.

1.4 Organization of the Thesis

In the following sections, the proposed framework for soft and hard fusion with examples and implementation details is discussed. Chapter 2 presents the fusion framework using Modified Dempster–Shafer Rule of Combination. In Chapter 3, methods of processsing data from AIS and Twitter are shown. Chapter 4 presents simulation results and discussions. Conclusions and future work are presented in Chapter 5.

1.5 Related Publications

- T.Abirami, Ehsan Taghavi, R.Tharmarasa, T. Kirubarajan and Anne-Claire Boury-Brisset , "Fusing Social Network Data with Hard Data, " Proc. of the Fusion conference, Washington DC, July 2015.

Chapter 2

Soft/Hard Data Fusion using Dempster–Shafer Theory

In this chapter the proposed fusion method to combine soft data from the social network with hard data from AIS is explained. In this thesis the Modified Dempster-Shafer rule of combination(MDRC) is used as an approach for event prediction and state estimation. It incorporates conflict measurement into Dempster rule of combination [27]. This work is motivated by the practical problem of detecting events from a finite place U, by combaining evidence concerning them, when evidence $A, B...etc$ about the events can be imprecise [15], [25], [49]. To understand the working of MDRC it is important to understand the concept of Dempster-Shafer theory and its basic operation.

2.1 Dempster–Shafer Theory

The Dempster-Shafer Theory can be interpreted as a theory of evidence and a theory of reasoning. It is a theory of evidence because it deals with weights of evidence and with numerical or boolean degrees of support based on evidence. It is a theory of reasoning because it focuses on the fundamental operation of reasoning: the combination of evidence. Assume there exists a set of n elemental propositions, called the frame of discernment(example: a_1, a_2, a_3 for $n = 3$). A proposition can be a hypothesis or a combination of hypotheses ($a_1 \cap a_2$ or $a_1 \cap a_2 \cap a_3$) [4]. These propositions can contain overlapping and even conflicting hypotheses [1, 41]. However, the frame of discernment, denoted by Φ, is a set of mutually exclusive and exhaustive propositions ($\Phi = \{a_1, a_2, a_3\}$). The power set can be defined as 2^Φ, which is the set of all subsets of Φ including the null set $\{\emptyset\}$ ($2^\Phi = \{\{a_1, a_2, a_3\}, \{a_1, a_2\}, \{a_2, a_3\}, \{a_1, a_3\}, \{a_1\}, \{a_2\}, \{a_3\}, \{\}\}$). Furthermore, the theory of evidence assigns a belief mass [10] to each element of the power set. The Belief Mass Assignment (BMA) assigns values $m(\xi)$ for all subsets $\xi \in 2^\Phi$ such that

$$m(\xi) \geq 0 \tag{2.1}$$

$$m(\emptyset) = 0 \tag{2.2}$$

$$\sum_{\xi \in 2^\Phi} m(\xi) = 1 \tag{2.3}$$

The belief mass that has been assigned to the proposition ξ can be treated as the certainty of the observer in the correctness of ξ. The actual state of the system is represented by the elements of the power set concerning its propositions, by containing all and only the states in which the proposition is true. The Dempster–Shafer theory

belief function is then the total belief that the proposition is true [29]. The belief function $b(\cdot)$ of a particular proposition ξ is given by

$$b(\xi) = \sum_{\theta \subseteq \xi} m(\theta), \quad \xi \in 2^{\Phi} \tag{2.4}$$

2.1.1 Uncertain Measurements

In general, the theory of evidences can be applicable to both crisp and fuzzy datasets [29]. In a evidence theory using crisp data, a hypotheis (eg: null hypotheis) can either be accepted or rejected, but by using fuzzy data a hypothesis can have certain degree of acceptability (eg: null hypothesis = μ, alternate hypothesis = $1 - \mu$, where $\mu \in [0, 1]$). An example of crisp hypothesis is "A person passed in the test" and an example of fuzzy hypothesis is "a person passed with good marks".

Raw data from sensors and social network sites comes with a lot of uncertainties and in order to deal with the uncertain measurements it is needed to use fuzzy belief mass assignment to assign values to the measurements. Assuming that giving BMA based on subsets ψ with associated values $m(\psi)$ on the set of all the finite subsets $\pounds$, that is

1. $m(\psi)$ is a function defined on all closed subsets $\psi \subseteq \pounds$
2. $m(\psi) \geq 0$ for all ψ
3. $m(\psi) \neq 0$ for only a finite number of ψ (which are focal subsets of m)

then the following is true :

$$\sum_{\psi \subseteq \pounds} m(\psi) = 1 \tag{2.5}$$

where the summation is well defined because of the third property. The function $m(\psi)$ is known as the Dempster–Shafer measurement. Let us consider z as an observation or evidence and ψ is the proposed hypothesis for the observation z. Let $\{\psi_1, \psi_2, \ldots, \psi_k\}$ be the focal subsets of m, where k is the last focal subset [13], [29]. One of the hypotheses constrains z to be in ψ_1, i.e., $z \in \psi_1$ with the wight associated it as $m(\psi_1)$. The other hypothesis can be constraining z to be in ψ_2, i.e., $z \in \psi_2$ with the wight $m(\psi_2)$ and so on [29]. If we know nothing about the measurement z, then the weight of the hypothesis associated with that is the total value of $m(\pounds)$ (the null hypothesis).

A fuzzy/vague measurement is an uncertain measurement whose focal subsets are linearly ordered under the set theoretic inclusion (nested). A Fuzzy Dempster–Shafer Belief Mass Assignment (FBMA) $m(\psi)$ is defined by the same properties as normal BMA [29] except now $m(\psi)$ is a fuzzy membership function which belongs to a fuzzy space χ. As such the following is true

$$\sum_{\psi \subseteq \chi} m(\psi) = 1 \tag{2.6}$$

The logical meaning of $m(\psi)$ can be given by the fact that each ψ is a fuzzy hypothesis about the observation of z [29]. Let $\{\psi_1, \ldots, \psi_k\}$ be the focal fuzzy subsets of m and assume that they are finite–level. It is unclear that z is or is not constrained by a particular subset $\psi_{1,1}$, therefore $\psi_{1,1}$ must be treated as initial guess about the meaning of the first fuzzy hypothesis ψ_1, where $\psi_{i,j}$ is j^{th} subset of i^{th} focal fuzzy subset ψ_i. By sequencing these into nested sequence $\{\psi_{1,1} \subseteq ... \subseteq \psi_{1,k}\}$, it can elaborate the nature of the uncertainty involved in the hypothesis ψ_1. This defines the finite–level fuzzy membership function ψ_1. Therefore one can interpret the remaining focal set

$\{\psi_2, \ldots, \psi_k\}$ in the same manner, where k is the last subset of m [29].

Consider that having FDS measurements m, m'. Then the FDS combination of m and m' is given by [28, pp. 144, Eq. (4.129)]

$$(m * m')(\psi'') = 0 \text{ if } \psi'' = 0 \tag{2.7}$$

and if $\psi'' \neq 0$, then

$$\begin{aligned}(m * m')(\psi'') &= \alpha_{\text{FDS}} (m * m')^{-1} \\ &\quad \times \sum_{\psi \cdot \psi' = \psi''} m(\psi) \cdot m(\psi')\end{aligned} \tag{2.8}$$

where $\alpha_{\text{FDS}}(m, m') \neq 0$ and the FDS agreement of m, m' is

$$\alpha_{\text{FDS}}(m * m') = 1 - \sum_{\psi \cdot \psi' = 0} m(\psi) \cdot m(\psi') \tag{2.9}$$

Here, $(\psi \cdot \psi') \triangleq \psi(z) \cdot \psi'(z)$ and the event $\psi \neq 0$ means $\psi(z) \neq 0$ for at least one z.

2.1.2 Drawbacks of Dempster Rule of Combination

The fusion rule $(m * m')(\psi'')$ is defined only if the agreement is non-zero because of the condition in which Eqn (2.9) is defined. So the factor α_{FDS} has an effect of completely ignoring the conflict and attributing any mass associated with the conflict to null set. For example let's consider that the FoD and the propositions are {A, B, C} and the hypothesis from two different sources are same as FoD. Assume that the first sources assigns mass to the hypotheses as $\{m(A) = 0.99, m(B) = 0.01\}$ and the second source assigns the mass as $\{m'(B) = 0.01, m'(C) = 0.99\}$. It is clear that

both the sources believes that the second hypothesis (B) is unlikely and they have conflicting views on the other two hypothesis (A) and (B). If we use Dempster's rule of combination to combine the evidences from two sources using Eqn (2.8), then we get

$$(m * m')(A) = \frac{0.99 \times 0}{1 - (0.99(0.01 + 0.99) + 0.01(0 + 0.99) + 0(0 + 0.01))} = 0, \quad (2.10)$$

$$(m * m')(B) = \frac{0.01 \times 0.01}{1 - (0.99(0.01 + 0.99) + 0.01(0 + 0.99) + 0(0 + 0.01))} = 1, \quad (2.11)$$

$$(m * m')(C) = \frac{0 \times 0.99}{1 - (0.99(0.01 + 0.99) + 0.01(0 + 0.99) + 0(0 + 0.01))} = 0. \quad (2.12)$$

This inference from this fusion rule is very counter intuitive, as it assigns the mass 1 for the proposition (B) that is believed by both sources as unlikely. In the context of this thesis, hard data and soft data are two different sources and it is very likely that these data may be conflicting with each other. So it is important that we have a Modified Dempster's Rule of Combination (MDRC) that takes conflicts in the data into the account.

2.1.3 Conflict Measurement

It has been identified that there may arise two types of conflicts in FBMA [21] and they are

- Discord - Disagreement in hypothesizing the propositions
- Non Specificity - Two or more propositions were not included in any hypothesis

These conflicts can be modeled based on pignistic probability distribution [21, 45].

$$CM(m.m') = -\sum_{\psi\in\chi}\sum_{\psi'\subseteq\chi} m(\psi')\frac{\rho(\psi\cap\psi')}{\rho(\psi')} \times log_2\left(\sum_{\psi'\subseteq\chi} m'(\psi')\frac{\rho(\psi\cap\psi')}{\rho(\psi')}\right) \quad (2.13)$$

Where the set function $\rho(-)$ is defined by $\rho(\omega) = 1$ if $\omega \neq \emptyset$ and $\rho(\omega) = 0$ otherwise. This type of Conflict Measurement (CM) is also called as pignistic entropy. The lower the value of CM is, the less conflict the sources has and higher the value of CM is, the more conflict the sources has. For the scenario explained in section 2.1.2, following is the conflict calculation.

$$CM(m.m') = -\left(0.99 \times log_2(0) + 0.01 \times log_2(0.01) + 0 \times log_2(0.99)\right) = \infty \quad (2.14)$$

$$CM(m'.m) = -\left(0 \times log_2(0.99) + 0.01 \times log_2(0.01) + 0.99 \times log_2(0)\right) = \infty \quad (2.15)$$

Conflict Measure for each data has the value of infinity and it means that the given data in section 2.1.2 has the most possible conflict in them. As we will be coding this in a programming platform, the data with zero values will be replaced by smallest possible value (10^{-20}) for avoiding error and indeterminate values.

2.1.4 Modified Dempster's Rule of Combination

Modified Dempster's rule that takes conflict measurements into the account is proposed as following [16].

$$m *_q m'(\psi) = \sum_{\psi'\cap\psi''=\psi} m(\psi')m'(\psi'') + \sum_{\psi\cap\phi=\emptyset} w_1 m(\psi)m'(\phi) + \sum_{\psi\cap\phi=\emptyset} w_2 m(\phi)m'(\psi). \quad (2.16)$$

where

$$w_1 = \frac{e^{-CM(m.m')}}{e^{-CM(m.m')} + e^{-CM(m'.m)}} \tag{2.17}$$

$$w_2 = 1 - w_1 \tag{2.18}$$

This approach as proposed, does not need the prior information and only uses the these weight function derived by comparing the evidence sources. Now when we apply this Modified Dempster's Rule of Combination to the example in Section 2.1.2, we get w_1 and w_2 as

$$w_1 = \lim_{x \to \infty} \frac{e^{-x}}{e^{-x} + e^{-x}} = 0.5 \tag{2.19}$$

$$w_2 = 1 - w_1 \qquad = 0.5 \tag{2.20}$$

Although the above equation will return indeterminate value (NaN) in a conventional mathematical calculation, the careful asymptotic analysis will reveal that its actual value approaches 0.5 as the exponential argument approaches infinity. Using this weights, we get the following calculations for combining evidences.

$$(m * m')(A) = 0.99 \times 0 + 0.5 \times 0.99(0.01 + 0.99) + 0.5 \times 0(0 + 0.01) \quad = 0.4950, \tag{2.21}$$

$$(m * m')(B) = 0.01 \times 0.01 + 0.5 \times 0.01(0 + 0.99) + 0.5 \times 0.01(0 + 0.99) = 0.0100, \tag{2.22}$$

$$(m * m')(C) = 0 \times 0.99 + 0.5 \times 0.99(0.99 + 0.01) + 0.5 \times 0(0.01 + 0) \quad = 0.4950 \tag{2.23}$$

This calculation is summarized in table 2.1. It is clear that MDRC takes the conflict into account and puts forth a reasonable combination of evidences.

Table 2.1: Advantage of MDRC over DRC

Hypothesis	**Source 1**	**Source 2**	**DRC fusion data**	**MDRC fusion data**
A	0.9900	0	0	0.4950
B	0.0100	0.0100	1.0000	0.0100
C	0	0.9900	0	0.4950

Chapter 3

AIS/ Twitter Data Processing

3.1 AIS Data Processing for Vessel Tracking

This section explains the statistical analysis of vessel motion patterns in the ports and waterways using the self reporting AIS data. As mentioned in introduction, maritime surveillance of ports and coastlines is very important, since there are lot of reported historical incidents like maritime terrorism, piracy, maritime pollution, unauthorised maritime arrivals, prohibited imports/exports etc. And most of the import and export operations of any country in the world is through the waterways. A potential terrorist attacks in ports or waterways could cause severe disruptions with major economic implications.

Generally, for maritime surveillance of ports and waterways, the sensors in use are radar, infrared and video cameras installed on fixed ground locations or mounted on border patrol vessels, aircraft and satellites. There are also a number of self-reporting systems available, which are mainly used for the navigation safety and collision avoidance. These messages transmitted by the self-reporting systems have thus become an

abundant and inexpensive source of information for maritime surveillance [12].

To improve the situational awareness in the maritime domain, it is important to find a best way to analyse the massive amount of broadcasted information from the AIS. By statistically analysing the AIS data, it is possible to detect the anomalies by detecting the deviation of a vessel from normal pathway. Using the historically available AIS training data, it is assumed that the relevant motion patterns have already been calculated during the data processing. In papers [5], [35], a system is described that uses the normal behavior of vessels to learn, detects anomalies and predicts the motion using an artificial neural network trained with historical AIS data. Tun *et al* [48] developed an algorithm, in which the vessel motion paths collected by an AIS receiver was broken down into seperate regions and it is based on the density maps.

To detect the anomalies in the vessel motion and to predict the vessel motion, AIS data is statistically analysed. Here it is assumed that the relevant motion patterns have been extracted from the historical data. This motion patterns can then be used in the algorithm to find the anomaly and to predict the motion. Using the framework adaptive kernel density estimation (KDE) [36], the anomaly detection process is carried out. Then it is applied squentially to the incoming AIS data. The probability of false alarm decides the threshold level of the anomaly detection. The algorithm for vessel motion prediction using the location and velocity within the specific time is derived under the assumption of null hypothesis(normal behaviour) and AIS data historic motion pattern. In a network of vessel paths, it is important to define the set of motion pattern origins, here it is $\Theta = \{1, 2, 3, 4\}$. The motion pattern P_j with origin $j \in \Theta$ consists of a trajectory, which can be denoted as

$$U_j^i = \{u_j^i(t); t = t_1, t_2, ..., t_{T_i}\} \tag{3.1}$$

where $i = 1, 2, ..., N_j$ is the index of the trajectory (N_j is the total number of training trajectories from P_j) and $u_j^i(t)$ is the kinematic state of the vessel traversing trajectory U_j^i at time t. Kinematic states of the vessel consists of the positional information (X, Y) and velocity information $(\dot{X}, \dot{Y})$ i.e. $S_j^i(t)$ is a four-dimensional vector $[X, Y, \dot{X}, \dot{Y}]^T$. It is assumed that kinematic states $u_j^i(t)$ are independent in t. It is important to note that, if $u_j^i(t)$ are the outputs of a tracking filter, then they are correlated in time.

By making use of all available trajectories from motion pattern P_j which serve as the training data set for normal behavior, this method can solve two types of problem. First, the problem of anomaly in motion detection. In this problem, the anomaly in the vessel motion is to be detected sequentially, if the state vectors of a test trajectory $(v_j(t) \in V_j)$ under the normal behavior (here normal behaviour is the null hypothesis). The second problem is motion prediction, in which the state of a test vessel at time t, $(v_i(t))$, act in accordance with the normal behaviour, it is required to predict the state of this vessel at time $t + T, (T > 0)$, under the assumption that it will continue to follow the pattern of normal behavior.

3.1.1 Anomaly Detection

It is possible to detect the anamoly in vessel motion using the training data from pattern P_j to determining a detection threshold that will partition the state space into two regions, one region to correspond to hypothesis H_0 (normal behavior), the other to H_1 (anomaly). This threshold is applied sequentially to the incoming test

data $v_j(t)$, this threshold is applied sequentially [36]. One-class classification and novelty detection [39] are some methods to solve this type of problem [18] using the support vector machines (SVM). Here the anomaly detection is performed using the adaptive kernel density estimator.

Notation index k is introduced to enumerate pairs (t, i) in (3.1), to simplify the equations. The set of unlabeled training data corresponding to motion pattern $j \in \Theta$ can then be written as,

$$\xi_j = u_{j,k}; k = 1, ..., K_j \tag{3.2}$$

where $K_j = \Sigma_{i=1}^{N_j} T_i.\xi_j$ is a random sample from the underlying multi-variate pattern probability detection function(pdf) under the null hypothesis $H_0 : p_j(u|H_0) = p_j(X, Y, \dot{X}, \dot{Y}|H_0)$. For motion anomaly detection, it is important to first approximate the density of vessel motion pattern to null hypothesis. This will be carried out using the adaptive KDE approximation [44].

3.1.2 Kernel Density Estimation

The density $s_j(x)$ is constructed by the KDE approximation by placing a kernel function ψ on every observed data $x_{j,k}$. The kernel is parameterized by its width h, which can be either fixed (identical for all observed data) or adaptive [36]. For simplicity, index j is dropped from notation in the reminder of the section. Let us assume $x \in \Re^d$, in this case $d = 4$ and the fixed KDE approximation is given by

$$s(x) \approx \tilde{s}(x) = \frac{1}{Kh^d} \sum_{k=1}^{K} \psi \left(\frac{x - x_k}{h} \right) \tag{3.3}$$

The kernel must satisfy $\psi(x) \geq 0$ and $\int_{\Re^d} \psi(x)dx = 1$. Adopting the Gaussian kernel with zero-mean and the covariance matrix Σ,

$$\psi(x) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right) \tag{3.4}$$

by using (3.4) in (3.3), we have

$$\tilde{s}(x) = \frac{1}{Kh^d(2\pi)^{d/2}\sqrt{|\Sigma|}} e^{-\frac{(x-x_k)^T\Sigma^{-1}(x-x_k)}{2h^2}} \tag{3.5}$$

The optimal fixed bandwidth, under the assumption that the underlaying pdf is Gaussian for the Gaussian kernel, is computed as

$$h^* = BK^{-\frac{1}{d+4}} \tag{3.6}$$

where

$$B = [4/(d+2)]^{1/(d+4)} \tag{3.7}$$

As a sample covariance, the covariance Σ needs to be estimated from the data. The fixed KDE is unable to deal satisfactorily with the tails of distributions since the observed data in the tails are rare, the window widths in the tails need to be broader. An obvious problem is deciding whether or not an observation is in a region of low density. The adaptive KDE approach [44] gets along with this by a two-stage procedure [36], where the first stage is typically the fixed KDE and the second stage computes adaptive window widths $\tilde{h}_k, k = 1, ..., K$ as $\tilde{h}_k = h * \lambda_k$ where h^* is already given and

$$\lambda_k = \left(\frac{\tilde{s}(x_k)}{l}\right)^{-\gamma} \tag{3.8}$$

where l is the geometric mean of a sequence $\tilde{s}(x_k)$ that is

$$\log l = \frac{1}{K}\sum_{k=1}^{k} log\tilde{s}(x_k), \tag{3.9}$$

and γ is the sensitivity parameter such that $0 \leq \gamma \leq$ and γ is typically set 0.5. Finally the adaptive equation is similar to the equation 3.3, except that for each x_k it is applied with the kernel of width $\tilde{h}(x_k)$

$$\tilde{s}(x_k) = \frac{1}{K}\sum_{K}^{k=1} \frac{1}{\tilde{h}_k^d}\psi\left(\frac{x - x_k}{\tilde{h}_k}\right) \tag{3.10}$$

3.1.3 Anomaly Detection using KDE

The computation in the state space of the decision boundary is very expensive and therefore to perform the detection using the values of density $g(x) = p(x|H_0)$ is proposed [36]. Let y be a test data from a test trajectory Y, which originates from the same node as the training data node j (subscript being suppressed). The anomaly, i.e. hypothesis, H1 is declared if

$$s(y) > \alpha s(x_r), \tag{3.11}$$

where

$$r = \arg\min s(x_k) \tag{3.12}$$

and $a > 0$ is a detector parameter which will be related to the probability of false detection. Note that if all training data $x_k; k = 1, ..., K$ falls inside the anomaly detection boundary, then $\alpha < 1$.

α is selected so that

$$P_r s(y) < \alpha g(X_r)|H_0 = P_{fa} \tag{3.13}$$

where P_{fa} is a specified probability of false alarm (incorrect anomaly detection). Then the above equation can be written as

$$\int \chi_{[0,\alpha s(x_r)]}(g(y))g(y)dy = P_{fa}, \tag{3.14}$$

where $\chi_A(z)$ is an index function defined as

$$\chi_A(z) = \begin{cases} 1, & z \in B \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

For a given P_{fa}, α can be computed numerically. A random sample $y_m \sim g(y) : m = 1, ...M$ is generated. Substitution of approximation $g(y) \approx \frac{1}{M}\Sigma_{m=1}^{M}\delta(y - y_m)$ into 3.14 yields,

$$P_{fa}(\alpha) \approx \frac{1}{M}\sum_{m=1}^{M}\chi_{[0,\alpha s(x_r)]}(g(y_m)) \tag{3.16}$$

We expect that $P_{fa}(\alpha)$ will depend on the number of training data points K.

3.1.4 Algorithm for (Anomaly) Vessel Motion Prediction

The motion prediction can be done by many available methods, to satisfy the high accuracy and quality in the output, the following method is to be selected. Variable Structure Interacting Multiple Model (VS-IMM) estimator is developed based on IMM algorithm and its effectiveness is shown in [23], [31] and [43]. However, models in the mode set are fixed in IMM but they can vary based on constraints in VS-IMM

[24].

Consider $S_{\kappa-1}$ is the segment in the effect, and the mode set is assumed as $U(\kappa-1) = (M_1(\kappa-1), M_2(\kappa-1), ..., M_i(\kappa-1), ...)$ in the time interval $(\kappa-2, \kappa-1)$. On the other hand, the segment in the effect is $S_{\kappa-1}$ during the time interval $(\kappa-1, \kappa)$ and $U(\kappa)$ is the IMM estimator mode set, which either include or remove the models from $U(\kappa-1)$. Using the sensors with low revisit rates the measurements can be obtained [24]. The measurements are received successively and the target changed its position from one segment to other upon the arrival of measurements. Let θ be the angle between the two segments, $\hat{x}(k-1|k-1)$ be the estimate at time step $k-1$ and $\hat{x}(k|k-1)$ be the predicted state at time step k. Figure 3.1 shows that, the target is in two different segments during the time interval of $(\kappa-1, \kappa)$. The time in which the target changes its segment is to be considered as $(\kappa-\tau)$, where the change is between $0 < \tau < 1$. By assuming the modes M_i, $M_{i'}$ and $M_{j'}$ were in effect during time interval $(\kappa-2, \kappa-1]T_1$, $(\kappa-1, \kappa-\tau]T_1$ and $(k-\tau, k]T_1$.

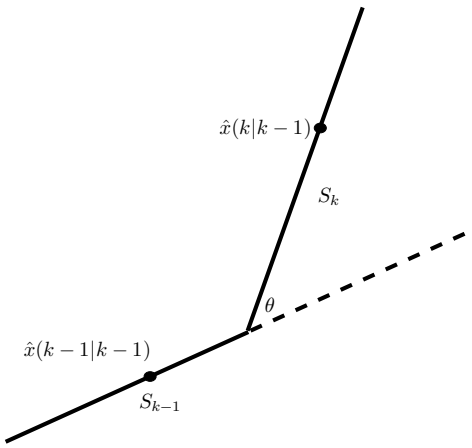

Figure 3.1: Changes in mode at time $\kappa-\tau$

Tracking algorithm steps:

Mixing probabilities calculation

Calculating mixing probabilities using,

$$\mu_{i|i'}(\kappa-1|\kappa-1) = \frac{1}{C_{i'}} P\{M_{i'}|M_i, Z_1^{\kappa-1}\}\ P\{M_i|Z_1^{\kappa-1}\} \tag{3.17}$$

The above equation for mixing probabilities can be rewritten as

$$\mu_{i|i'}(\kappa-1|\kappa-1) = \frac{1}{C_{i'}}\ [p_{ii'}(S_{\kappa-1})]\ \mu_i(\kappa-1) \qquad \forall i,\ \forall i' \tag{3.18}$$

where $p_{ii'}(S_{k-1})$ is the mode transition probability in segment S_{k-1}, $\mu_i(k-1)$ is the mode probability at time step $k-1$ and $C_{i'}$ is the normalizing constant

IMM Mixing

IMM mixing procedure is given by

$$\hat{x}^{0i'}(\kappa-1|\kappa-1) = \sum_{M_i \in U(\kappa-1)} \hat{x}^i(\kappa-1|\kappa-1)\ \mu_{i|i'}(\kappa-1|\kappa-1) \qquad \forall i' \tag{3.19}$$

$$\begin{aligned} P^{0i'}(\kappa-1|\kappa-1) = \sum_{M_i \in U(\kappa-1)} & \mu_{i|i'}(\kappa-1|\kappa-1)\ \Big\{\hat{x}^i(\kappa-1|\kappa-1) \\ & + [\hat{x}^i(\kappa-1|\kappa-1) - \hat{x}^{0i'}(\kappa-1|\kappa-1)] \\ & [\hat{x}^i(\kappa-1|\kappa-1) - \hat{x}^{0i'}(\kappa-1|\kappa-1)]^T\Big\} \qquad \forall i' \end{aligned} \tag{3.20}$$

Mode Matched Filtering

Mode matched filtering is given by

$$\hat{x}^{i'j'}(k|k-1) = F^{j'}(k, k-\tau)\ R\ F^{i'}(k-\tau, k-1)\hat{x}^{0i'}(k-1|k-1) \tag{3.21}$$

$$P^{i'j'}(k|k-1) = F^{j'}(k, k-\tau)\ R\ \Big\{F^{i'}(k-\tau, k-1)P^{0i'}(k-1|k-1)\ F^{i'}(k-\tau, k-1)^T \tag{3.22}$$

$$+Q^{i'}(k-\tau, k-1)\Big\}\ R^T\ F^{j'}(k, k-\tau)^T + Q^{j'}(k, k-\tau)$$

Updating Mode probability

The mode probability is updated as follows

$$\mu_{i'j'}(\kappa|\kappa) = \frac{1}{C_{i'j'}} \Lambda_{i'j'}(\kappa)[p_{i'j'}(S_{\kappa-1}, S_\kappa)]\mu_{i'}(\kappa - \tau) \tag{3.23}$$

where, $C_{i'j'}$ is a normalizing constant and $\mu_{i'}(\kappa - \tau)$ is the mode probability of model $M_{i'}$.

Combination of state estimate and covariance

The final state is given by

$$\hat{x}(\kappa|\kappa) = \sum_{M_{j'}, M_{i'} \in U(\kappa)} \hat{x}^{i'j'}(\kappa|\kappa)\ \mu_{i'j'}(\kappa|\kappa) \tag{3.24}$$

$$P(\kappa|\kappa) = \sum_{M_{j'}, M_{i'} \in U(\kappa)} \mu_{i'j'}(\kappa|v)\ \left\{ P^{i'j'}(\kappa|\kappa) + [\hat{x}^{i'j'}(\kappa|\kappa) - \hat{x}(\kappa|\kappa)]\ [\hat{x}^{i'j'}(\kappa|\kappa) - \hat{x}(\kappa|\kappa)]^T \right\} \tag{3.25}$$

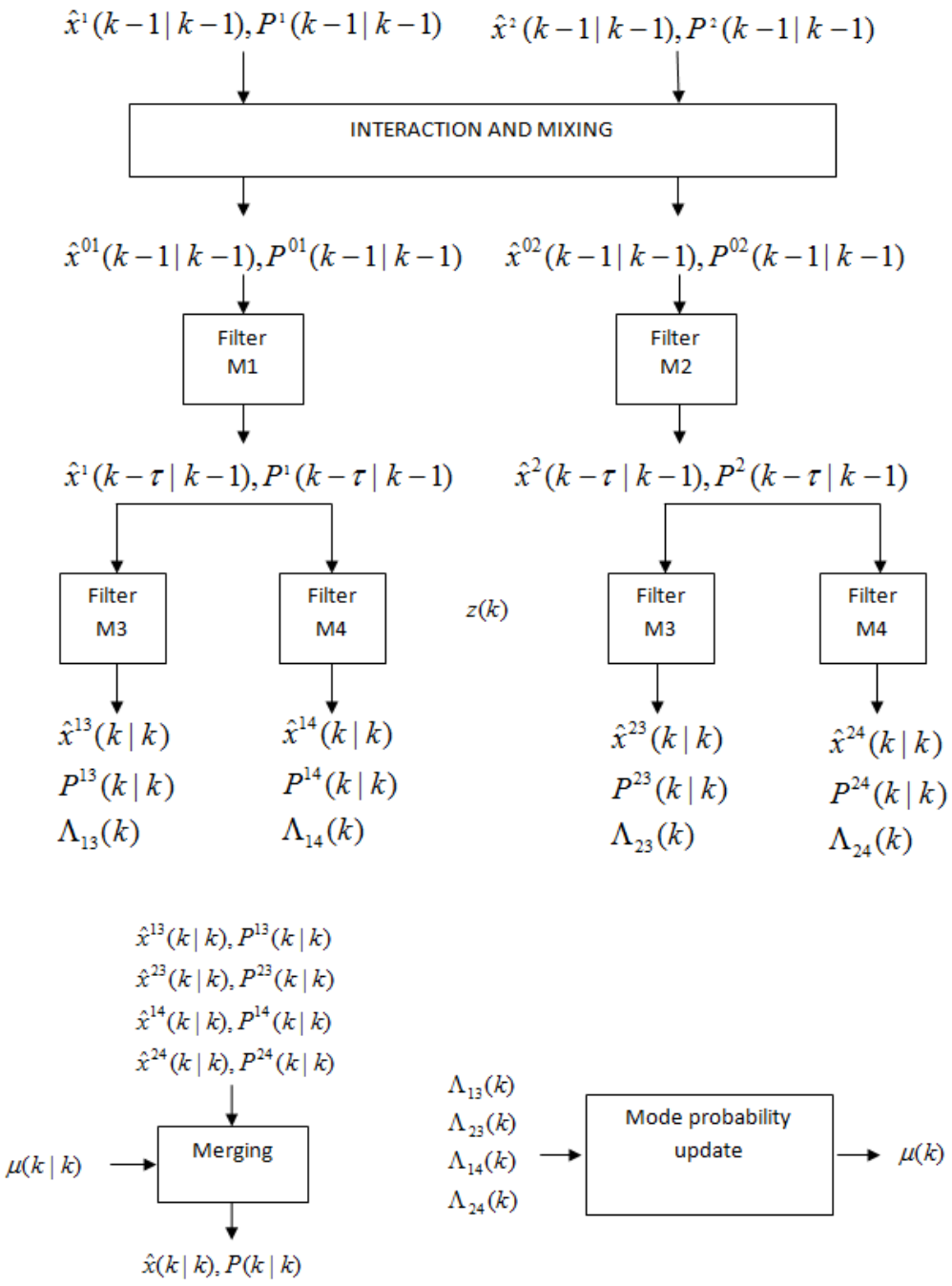

Figure 3.2: Filtering one cycle (Courtesy of [24])

3.2 Twitter data processing

In the recent times, there have been many man–made and natural disasters happening in the world. For example, man–made disasters like twin tower attack in USA, Taj hotel attack in India, and many revolutions that happened around the world killed thousands of people. On the otherhand, natural disasters like earthquake, tsunami, hurricane destroyed properties and killed the lives of people. Both type of disasters can be natural or man-made come without warning and cause massive damage to life and wealth of the people [2]. Nowadays, online social networking platforms such as Facebook, Twitter and YouTube have been playing an massive role in breaking news about the disasters to the people. People share information about the incidents and ask for help through social media, As there are millions of people who use such social networking sites in recent years. Many studies reveal that, the social networking sites play a major role in spreading the news and helping people during the man–made or natural disasters and played pivotal role in solving many problems in these types of disaster [46], [32].

Social networking sites will be active and helpful 24/7 to the people , even during the disasters. In general, the other conventional method which are in use for the communication will be useless during the disaster time. Thus, the social networking sites are considered to be a live monitoring sensor systems with varoius active sensors reporting the events through these sites [38], [37]. People in need of help can communicate their requirement through social media and social networking sites. When a disaster happens, real people share the knowledge, to create the situational awareness, to get the emergency help and assessment, voluntarily and involuntarily through socila networking sites. [20].

During the time of intense difficulty crisis coordination is important. Especially during the disasters, crisis coordination reduce the hazards and there are many organisations and individuals that held during crisis coordination. In large scale scale disasters, it is difficult to recover and restore to the normal life without the critical understanding about the disasters [19].

3.2.1 The problem in extracting data

Only in recent years, many general community started to use the social media as a new source of resource for the information. The conventional methods to gather the information about the incidents are supplimented by these information from social networking sites. Thus, it helps to create the situational awareness and helps restoring the safety. Watch officers are used to analyze the information by scanning and assessing the available data in a short period of time. It is clear that, the social networking sites will provide a rich source of information at free of cost. But, it is very difficult to convert this social networking sites large stream data as a data with useful information data for the situational awareness. To address this issue, some special tools and services are required to mine the social networking sites data and provide the following needs:

1. Unexpected or unusual incidents should be detected.

2. Create a summarised message about the incident by maintaining the awareness about the incident. By this , it is possible to eliminate reading all individual messages.

3. Understanding the impact of the incident on people and classify the high value

messages during the incident. (e.g. informations which tells about the damage and about the people who needs help)

4. During an incident, it is important to identify, track and manage the issues within the incident. This may last for days, weeks or months.

5. Perform forensic analysis of incidents by analysing social networking sites content from before, during, and after an incident.

In figure 3.3, the general infrastructure of capturing tweets and processing through various blocks are shown. By using the file and database writing services, it is possible to capture the tweeets and to store it for the processing. This data is more versatile than the available metadata from the Twitter API. The analysis and processing of tweets can be done using the location mapping services, burst detcting and clustering services. The messaging service is a programming method to communicate between the twitter capture virtual machines, file writer, database writer, burst detector and location mapping service. Watch officers will use the burst detection model at a perception level to create the awareness.

3.2.2 Detecting Incidents

The historical data is used to build a statistical model of word occurences. To detect the incidents, the burst detection method analyse the content of the tweets to detect the words which are of interest to the surviellance agencies. If there is any positive variation from the statistical model, then it is defines a s burst. To continuously monitor the incoming messages from the data capture virtual machines, this burst detection method is to be coupled with the complex event processing system.

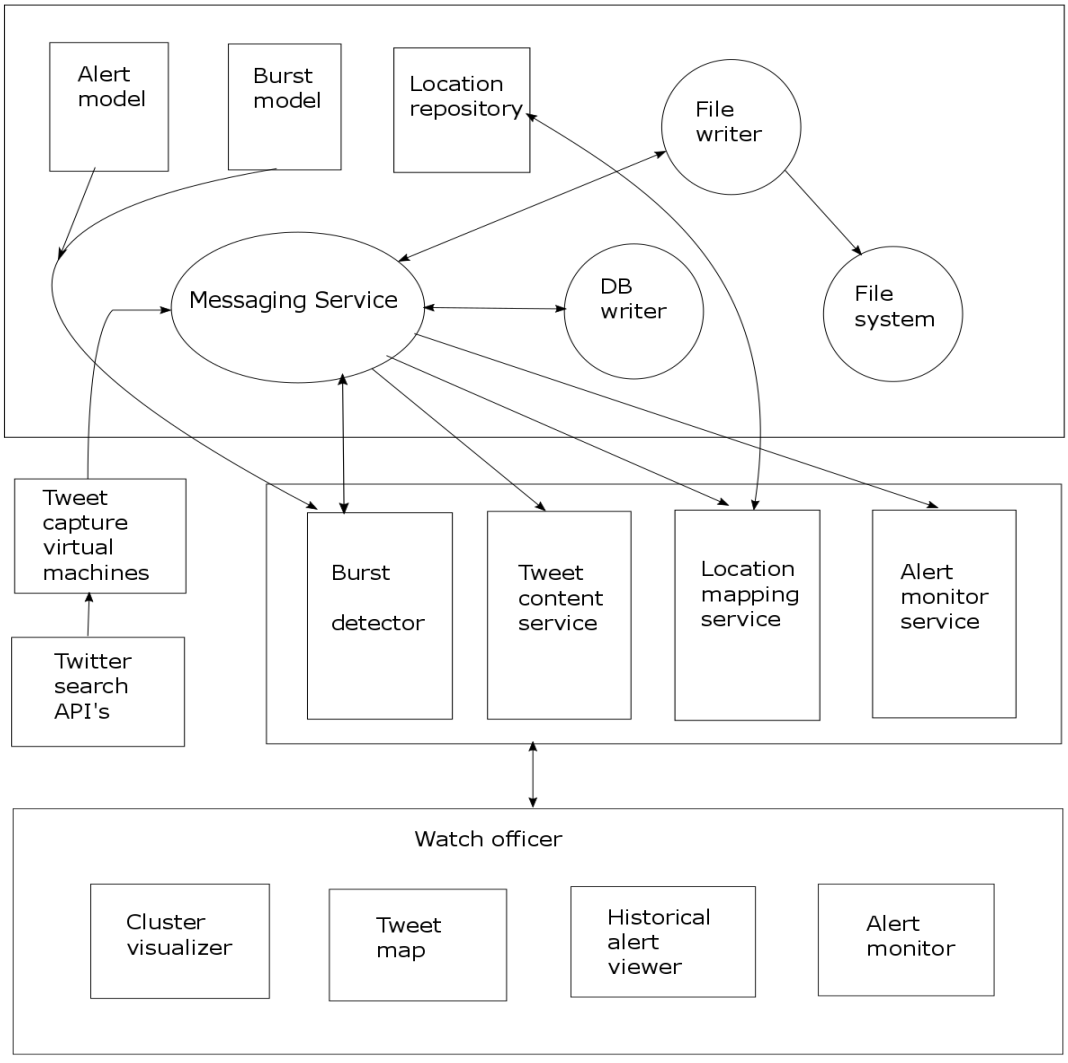

Figure 3.3: Twitter data processing

Chapter 4

Simulation Studies and Results

To validate the proposed algorithms, simulations are carried out for two different scenarios and results are analyzed. The simulated scenario based tests and results are presented in section 4.1 and 4.2.

Marine security is a growing concern [9]. There are various types of threats that can enter a country through its waterways. For instance, a small water craft can be turned into a weapon to destroy properties of a navy or to pirate a ship. Increased surveillance of sea is needed to protect countries from these types of threats. Presently, AIS signal and anomaly detection algorithms are used to filter the unusual behavior in maritime security systems. There are various types of sensors that are engaged in maritime security such as high frequency radar, active and passive sonar, and synthetic aperture radar. For example, Canada's CP-140 Aurora platform consists of many sensors that are used to collect data for the surveillance of surface targets [42].

In airborne surveillance of surface targets, sensor performance and data fusion are two major research areas. In this paper a novel approach is proposed for using the social network data (soft data) along with AIS or other airborne surveillance data

sensors data (hard data) by fusing them to get a better estimation for tracking surface targets and also to increase the maritime security. In order to understand how the proposed algorithm works under different circumstances, we investigate two different scenarios in the following subsections.

4.1 Scenario 1

This scenario assumes that the surveillance agencies know that a region will certainly be under attack but it requires help from Airborne or Marine sensors and Social Network (SN) to predict the exact place of attack. An unfortunate example of this case is 26/11 Mumbai terror attacks. In this case, the policing agency knew that the attack is underway in Mumbai region but was unable to pinpoint the exact locations for which the counter forces had to be dispatched [34]. There are various studies that shows that this attack can be tracked by SN data [8] (soft data) and Airborne/Marine surveillance data [6] (hard data) and this thesis proposes an outline for a tool that can fuse these data. Let's assume that the soft data is collected from a social networking website i.e.,Twitter. Figure 4.1 shows the tweets in an area with green and red dots. Also, consider that the red dots are the data which is filtered based on keywords. Each filtered data can then be sorted based on the location it specifies.

Let's assume a hypothetical situation where a region with 3 cities is under attack. Since the situation that is being modeled (terror attack) is certain, the sum of Belief Mass (BM) assigned to the cities is equal to one (i.e, propositions are city 1, city 2, city 3). Soft data collected from sources such as twitter can be filtered based on keywords such as "attack", "terrorists", "casualty" etc. and sorted based on location it specifies (in our hypothetical case keywords will be "city 1", "city 2" and "city 3").

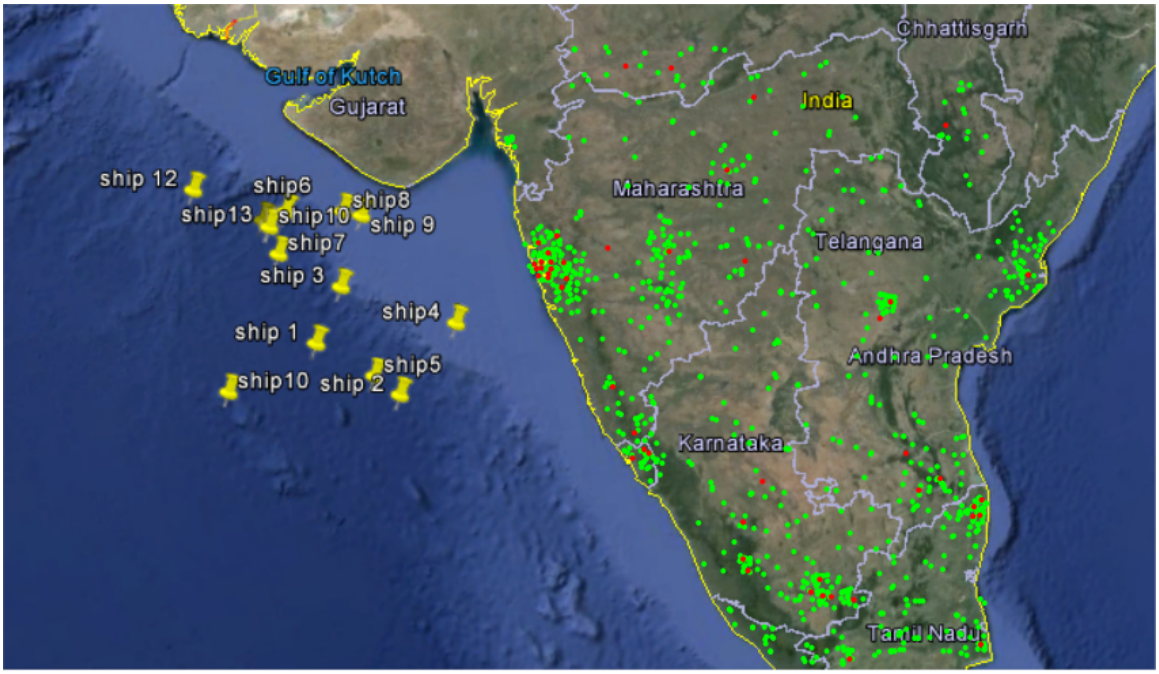

Figure 4.1: Ship movements and tweets in Mumbai area (courtesy of Google Earth)

BM of SN data can then be assigned based on simple formulations such as

$$BM_{SN}^{City_n} = \frac{\text{Number of filtered tweets about } City_n}{\text{Total number of filtered tweets}}, \tag{4.1}$$

where $n = 1, 2, 3$ in our hypothetical case. For this scenario a sample text file with list of hypothetical tweets was created and filtering and counting algorithms were employed to obtain the belief mass.

Table 4.1: Belief Mass assignment for SN data

Cities	**Tweets**	BM_{SN}
1	25	0.2380
2	43	0.4095
3	37	0.3523

On the other hand, let's assume that we have AIS data as hard data and we can employ Anomaly Detection algorithm as explained in Section 3.1 to predict the pathway of the anomalous vessel movement. Figure 4.2 shows the working of the tracking algorithm and anomaly detection. We can work out the BM for AIS data using angle of approach of anomalous vessel as shown in Figure. 4.3. If θ_n is the angle at which $city_n$ is located from the angle of approach of anomalous vessel, then BM can be calculated as

$$BM_{AIS}^{City_n} = \frac{\frac{\pi}{2} - \theta_n}{\sum_i \left(\frac{\pi}{2} - \theta_i\right)}, \tag{4.2}$$

where $n = 1, 2, 3$, in our hypothetical case. Using this assumption, belief mass for 3 cities as shown in Figure 4.3 can be calculated as shown in table 4.2.

Now that we have calculated the belief mass for both SN and AIS data, we can

Table 4.2: Belief Mass assignment for AIS data

Cities	**Angle (θ Radians)**	BM_{AIS}
1	0.3490	0.4001
2	0.5235	0.3428
3	0.7852	0.2580

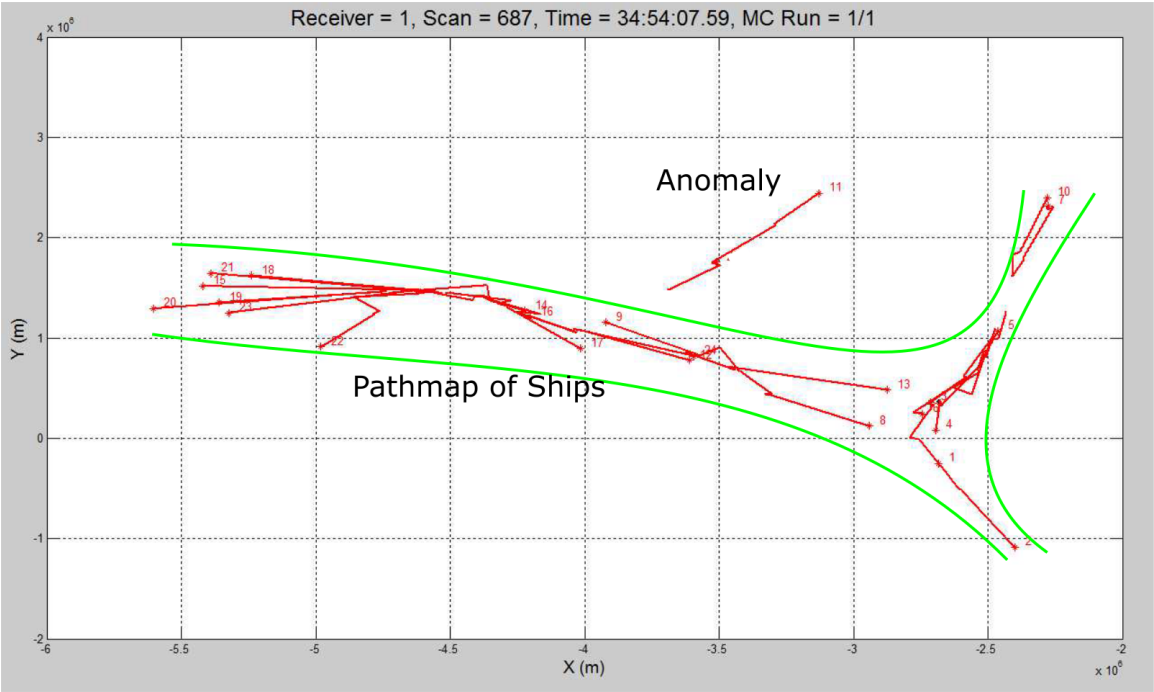

Figure 4.2: An example of tracking using AIS data and anomaly detection

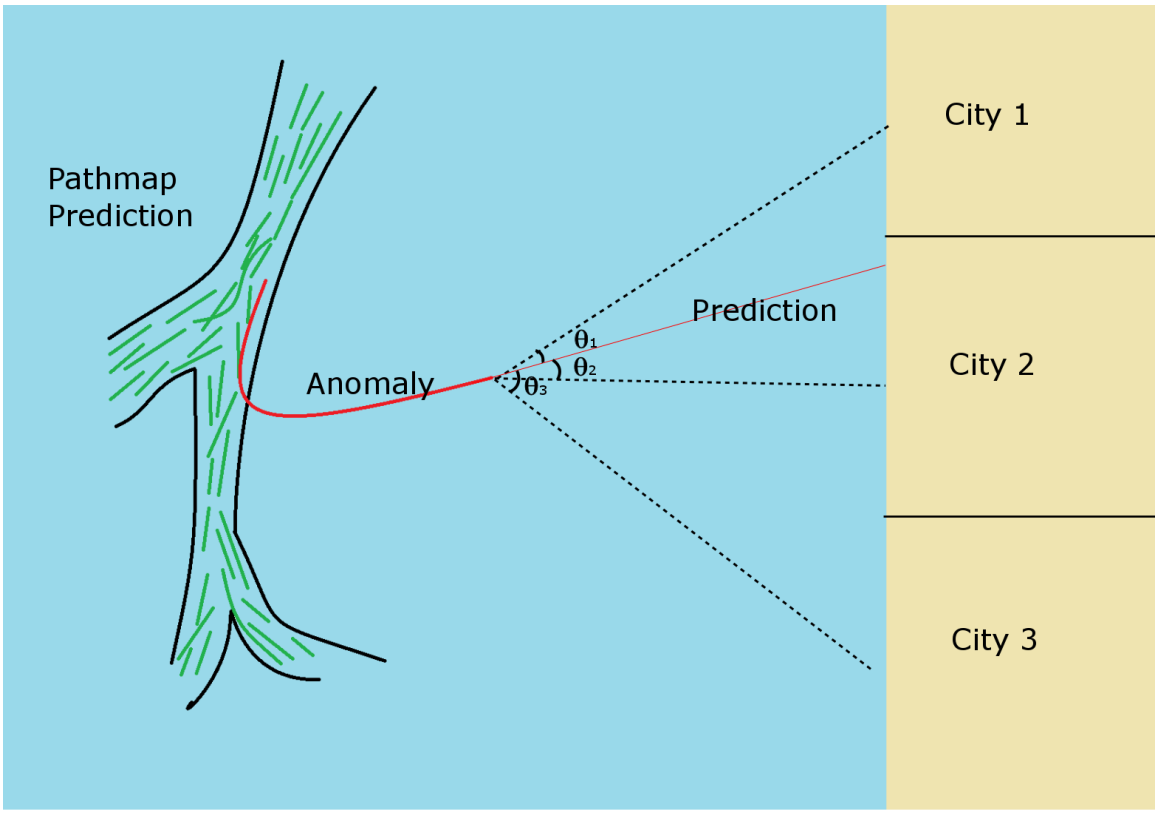

Figure 4.3: Approach of anomalous vessel and its predicted path

use DRC and MDRC algorithm as explained in Section 2 to fuse them. Fused data using DRC and MDRC method for the data given in Table. 4.1 and Table. 4.2 is shown in Table. 4.3. As shown, even in data with mild conflicts, fusion using DRC and MDRC shows significant difference. This difference is quite pronounced in data with visible conflicts as shown in Table. 4.4.

Table 4.3: Data fusion using DRC and MDRC in mild conflict situation

Cities	**AIS data**	**SN data**	**DRC Fusion**	**MDRC Fusion**
1	0.4001	0.2380	0.2923	0.3327
2	0.3428	0.4095	0.4309	0.3762
3	0.2580	0.3523	0.2790	0.3193

Table 4.4: Data fusion using DRC and MDRC in strong conflict situation

Cities	**AIS data**	**SN data**	**DRC Fusion**	**MDRC Fusion**
1	1.0000	0.3333	1.0000	0.5946
2	0	0.3333	0	0.2027
3	0	0.3334	0	0.2027

The tables in this section shows the fusion of discrete data sets but fusion can also be carried out for real time or continuous data as shown in Figure 4.4. It can be seen from this figure that DRC fusion gives unreasonable results when one of the data approaches either 1 or 0, especially in the plot for city 2, the DRC fusion closely follows the SN data and it appears that as if fusion doesn't take AIS data into account. On the other hand, MDRC approach takes a reasonable stand between SN and AIS data in all three cities.

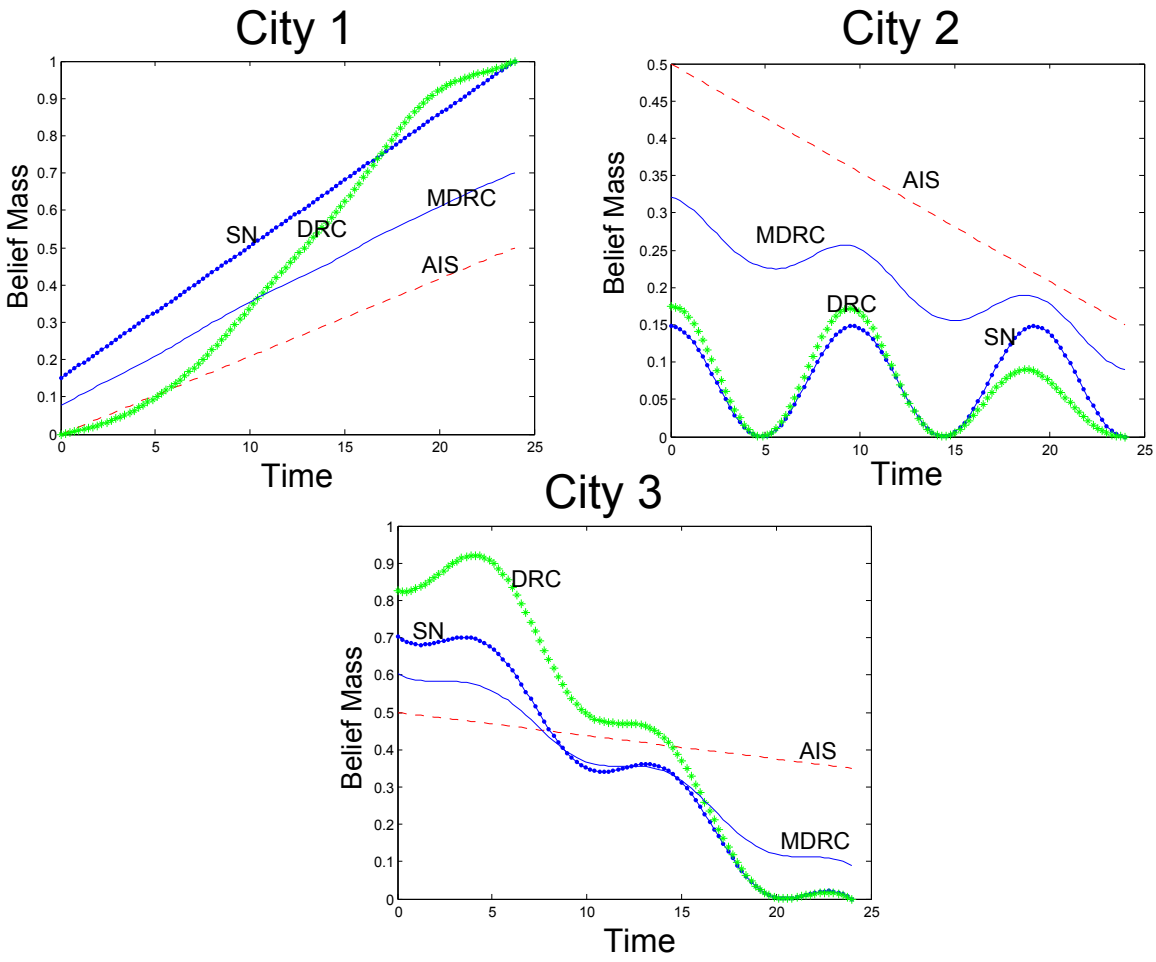

Figure 4.4: An example plot of real-time or continuous data fusion in scenario 1

4.2 Scenario 2

This scenario is similar to a classical threat monitoring system, where by fusing the scoured data from social netwrok space and sensor equipments, we might possibly predict a threat to a community. Scenarios such as community unrest, community clashes, riots etc., can be best predicted by publicly available social media information [26], whereas the external threats and cross-border terrorism can be best predicted using sensor hard data such as AIS [47]. There are also many studies that point towards the fact that the radical elements of the society are increasingly using social media to show their presence [50]. The best and unfortunate example of this is the Charlie-Hebdo attack and the subsequent murder of 12 of its employees, where the radical elements have been continuously threatening the Charlie-Hebdo establishment using social media.

This study explores the possibility to use both hard data (AIS as an example here) and soft data (Twitter as an example here) to gain a reasonable insight of threat to the community.

As an example of hard data, we use AIS maritime surveillance system. The AIS data provides all details such as the name, latitude, longitude, Speed Over Ground (SOG), Course Over Ground (COG), base station and destination of all the vessels in a particular maritime region. By applying the anomaly detection algorithms to AIS data the details of the anomalous vessels and the Probability of False Alarm (PFA) can be obtained. Using PFA and the angle at which the city is located to its predicted path, one can obtain the BMA for AIS data as

$$BMA_{AIS} = (1 - PFA) \times \frac{\pi - \theta}{\pi}. \tag{4.3}$$

For the same three city case as shown in Figure 4.3, the BMA assignment can be tabulated as shown in table 4.5 for PFA = 0.2110. Note that we are only trying to predict a threat and hence the threat is uncertain so the sum of the belief masses does not have to be one (In this case the propositions are "True", "False").

Table 4.5: Belief Mass assignment for AIS data

Cities	**Angle (θ Radians)**	BM_{AIS}
1	0.3490	0.7013
2	0.5235	0.6575
3	0.7852	0.5917

SN data is taken from twitter and the tweets are continuously monitored and observed for deviation from the statistical model build using past history. This burst detection mechanism is explained in more detail in section 3.2. A sample statistical deviation for synthetic data is shown in Figure. 4.5 for one particular keyword "Jihad". If we consider this as keyword i, the probability assignment (PA) can now be defined as

$$PA_i = \begin{cases} \frac{dev_t^i}{dev_\mu^i} & \text{if } \mu_{currentdata} > \text{Threshold} \\ 0 & \text{otherwise} \end{cases}, \tag{4.4}$$

where dev_t^i is deviation of mean of current data from the threshold and dev_μ^i is deviation of mean of current data from historical data. Both these terms are schematically presented in Figure 4.5. Note that this is a simple, but efficient formulation for PA only using one statistic (mean), but there can be many formulations based on multiple statistics such as mean, standard deviation, degree of skewness etc. BM for

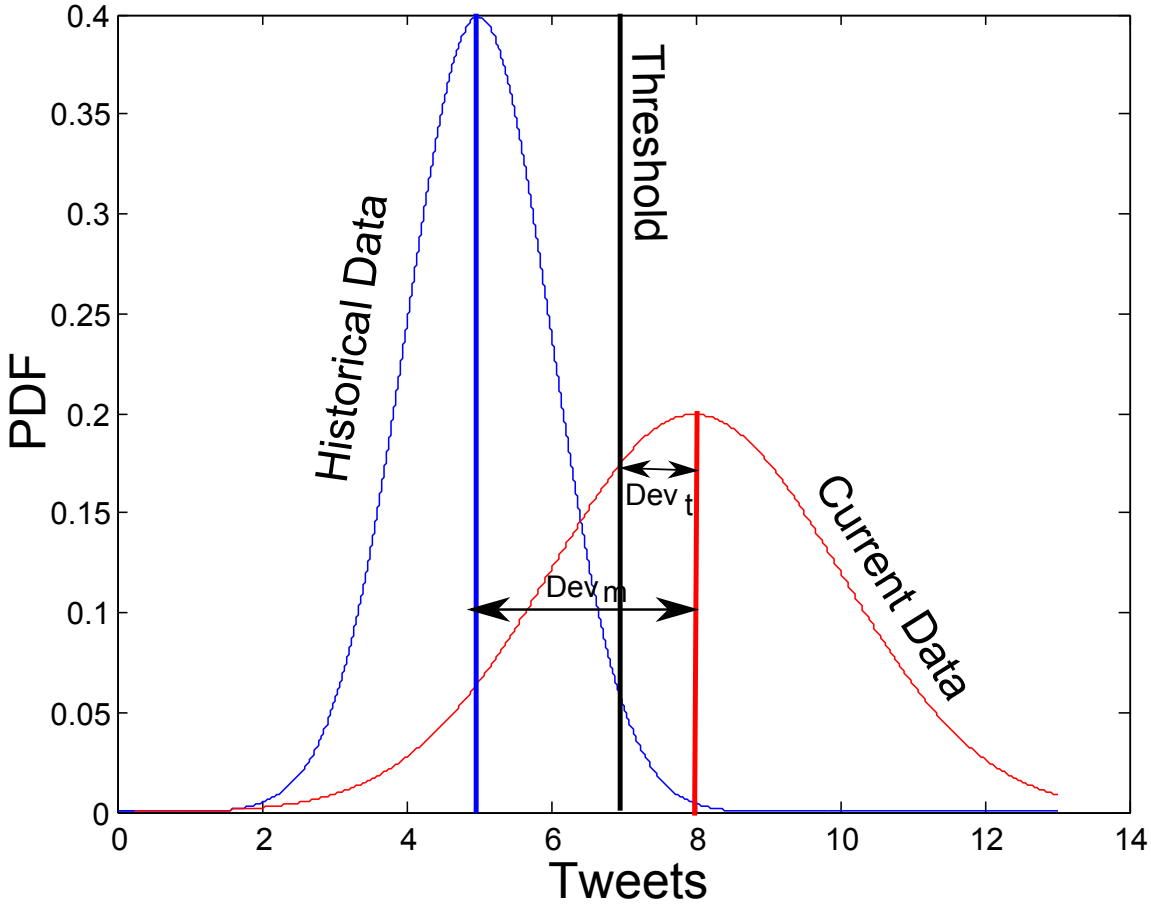

Figure 4.5: An example of statistical deviation in historical and current twitter data.

SN data can then be assigned as

$$BM_{SN} = \max_i PA_i \tag{4.5}$$

BM_{SN} calculated for the case shown in Figure 4.5 is 0.25. Since this data was not sorted on the basis of location, we assign the belief mass for all three hypothetical cities to be 0.25. With these SN and AIS data, the fusion using DRC and MDRC yielded the results as shown in Table 4.6. The difference between DRC and MDRC is significant and it is pronounced when one of the data points approaches the extreme value. The real disadvantage of DRC is visible in conflict situations as shown in Table 4.7.

Table 4.6: Data fusion using DRC and MDRC - Scenario 2

Cities	**AIS data**	**SN data**	**DRC Fusion**	**MDRC Fusion**
1	0.7013	0.2500	0.4390	0.4780
2	0.6575	0.2500	0.3902	0.4574
3	0.5917	0.2500	0.3257	0.4252

Table 4.7: Data fusion using DRC and MDRC in conflict situation - Scenario 2

Data	**AIS data**	**SN data**	**DRC Fusion**	**MDRC Fusion**
1	0	0.9999	0	0.5000
2	0.0001	1.0000	1.0000	0.5000
3	0	1.0000	NaN	0.5000

The comparison between the behavior of DRC and MDRC fusion can be explained using Figure 4.6. This figure shows DRC and MDRC fusion of synthetic 2000 data couples. In a data couple one is named as SN data and the other is named as AIS

data. DRC and MDRC fusion is carried out between SN and AIS data and SN and fused data is plotted against AIS data. It can be clearly seen that DRC fusion skews the fused data to the extreme value when AIS data has value near zero or one. But on the other hand, MDRC fusion squeezes the fusion around AIS data and we can expect that it will make a reasonable prediction.

These fusion algorithm can also be applied for continuous synthetic data or real time data. Following are the 5 cases of continuous data where we examine the behavior of DRC and MDRC fusion algorithm for different data trends.

CASE 1: In this case two oscillating data with different wavelengths are fused with DRC and MDRC and shown in Figure 4.7. Both fusion seem to work well as there are no values near zero or one. However when both SN and AIS data exceeds the value 0.6 then the DRC fusion give a result that is significantly higher but the MDRC fusion seem to take the middle ground between two data.

CASE 2: In this case a linearly increasing data and a linearly decreasing data is fused together and results are shown in Figure 4.8. In this case the DRC fusion has unpredictable behavior on the left side when the parent data is 0 and 0.9999. On the other hand MDRC fusion seem to have a predictable behavior.

CASE 3: In this case, a oscillating function and linear function is fused and shown in Figure 4.9. Both fused results has adopted the characteristics of parent data, but the MDRC fusion seem have a reasonable result at extreme parent values.

CASE 4: In this case exponentially increasing and exponentially decreasing functions are fused and result is shown in Figure 4.10. Similar to case 2 the behavior of DRC fusion seem to be erratic when parent values approached 0.001 and 1. On the other hand the MDRC fusion seem to have predictable behavior.

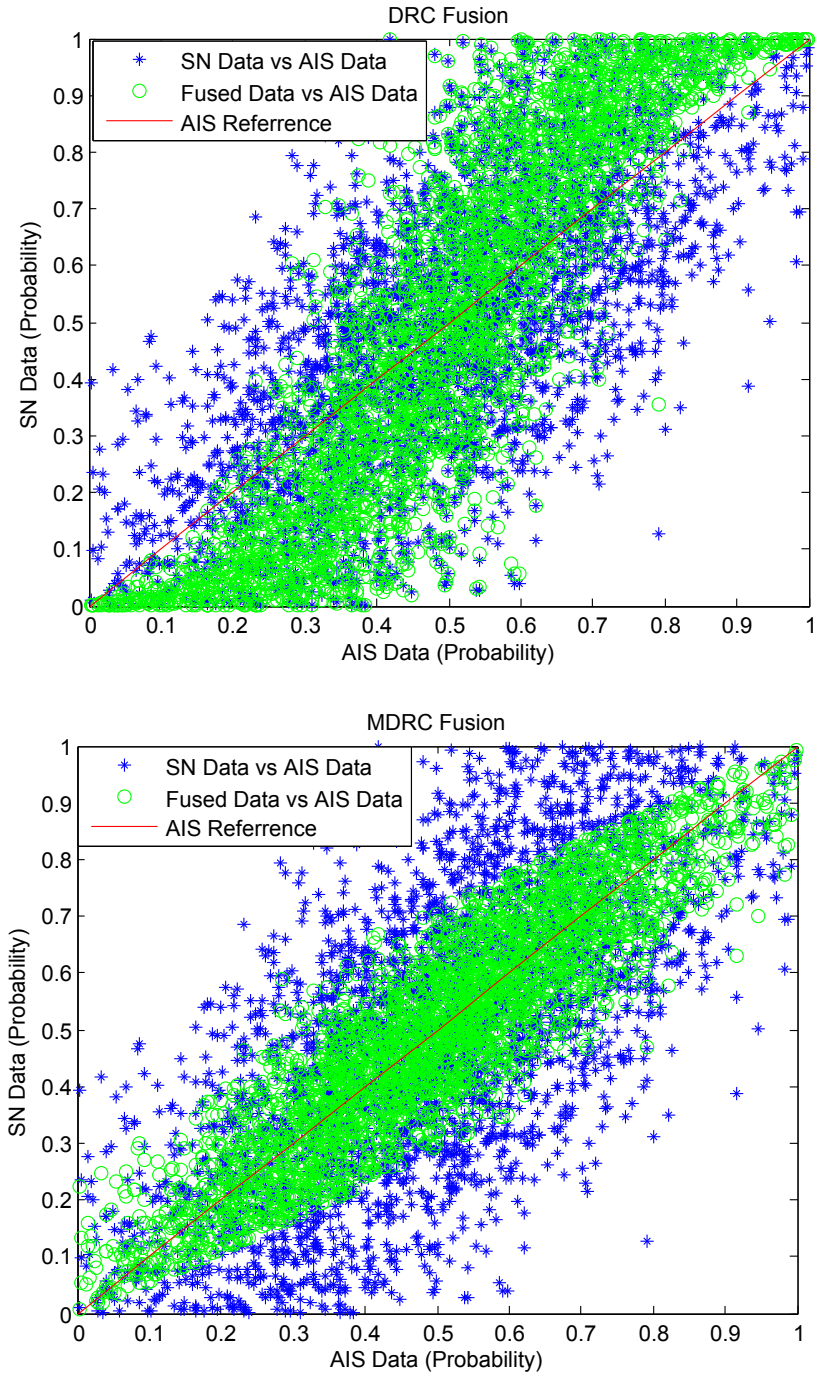

Figure 4.6: Comparison of DRC and MDRC fusion behaviour

CASE 5: In this case both data is linearly decreasing but one has a random noise incorporated in it as shown in Figure 4.11. Both fusion result had noise in them and gave reasonable results.

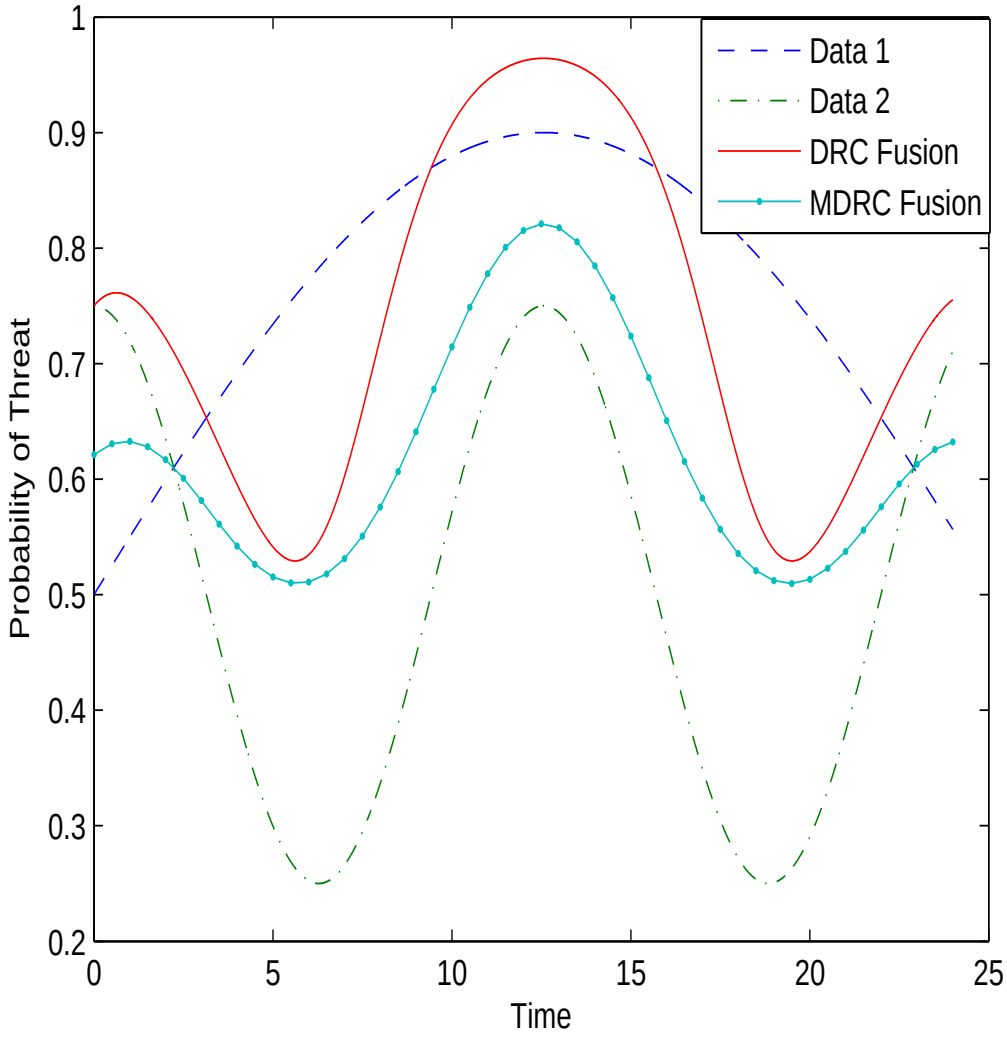

Figure 4.7: Data fusion in scenario 2 - case 1

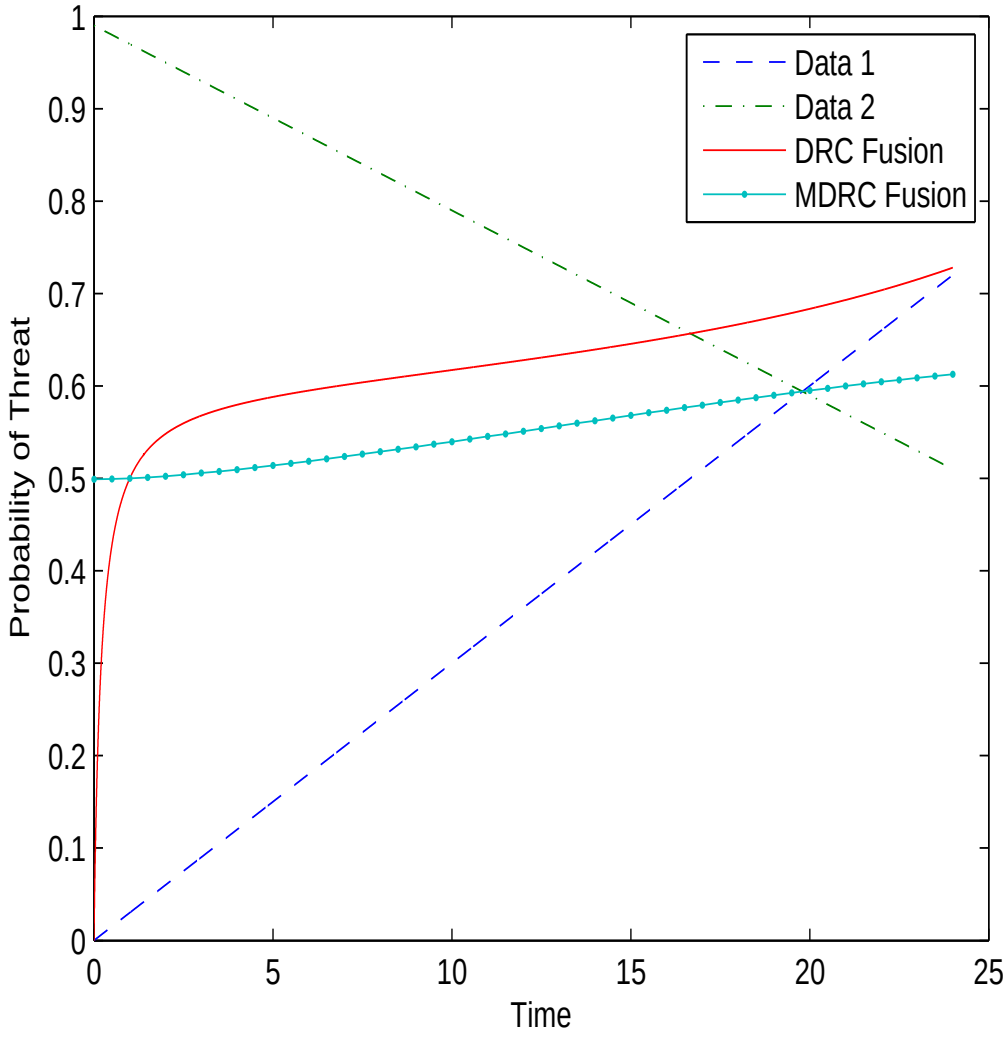

Figure 4.8: Data fusion scenario 2 - case 2

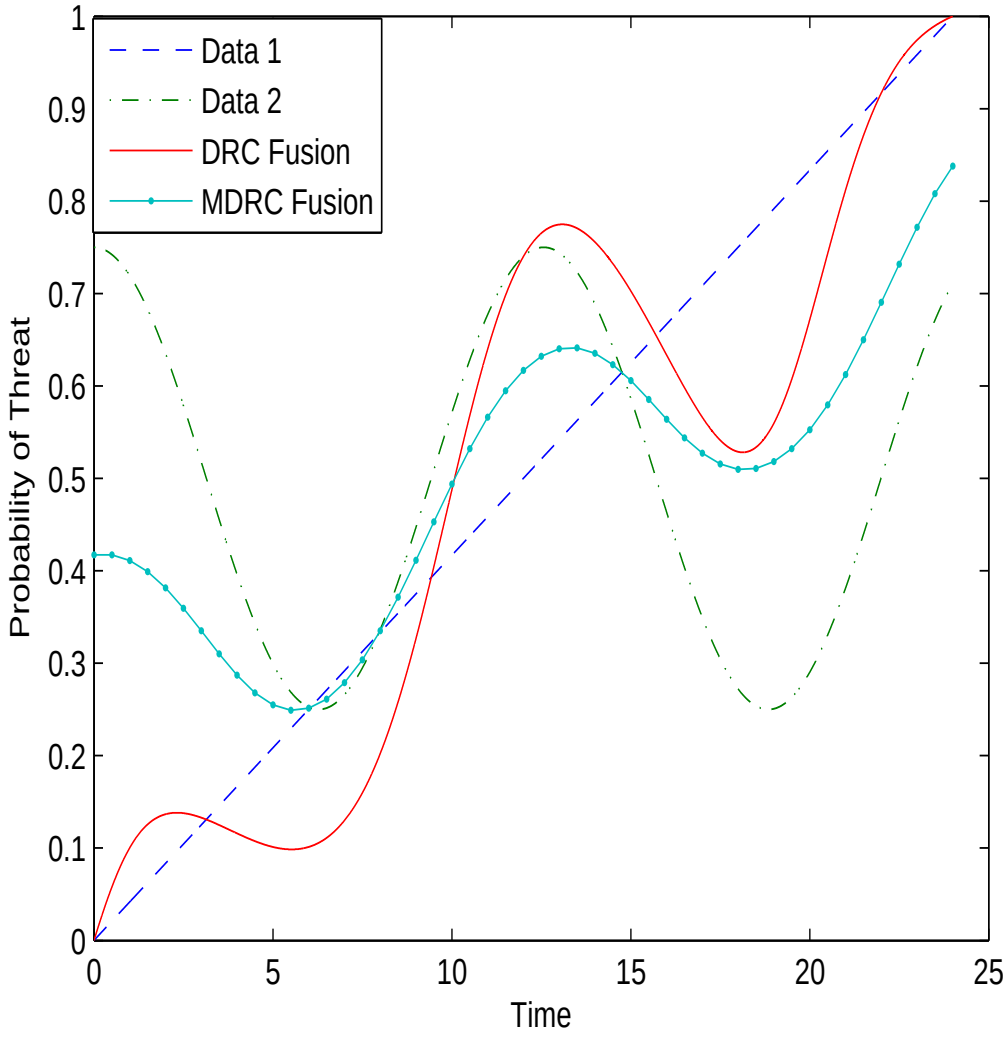

Figure 4.9: Data fusion scenario 2 - case 3

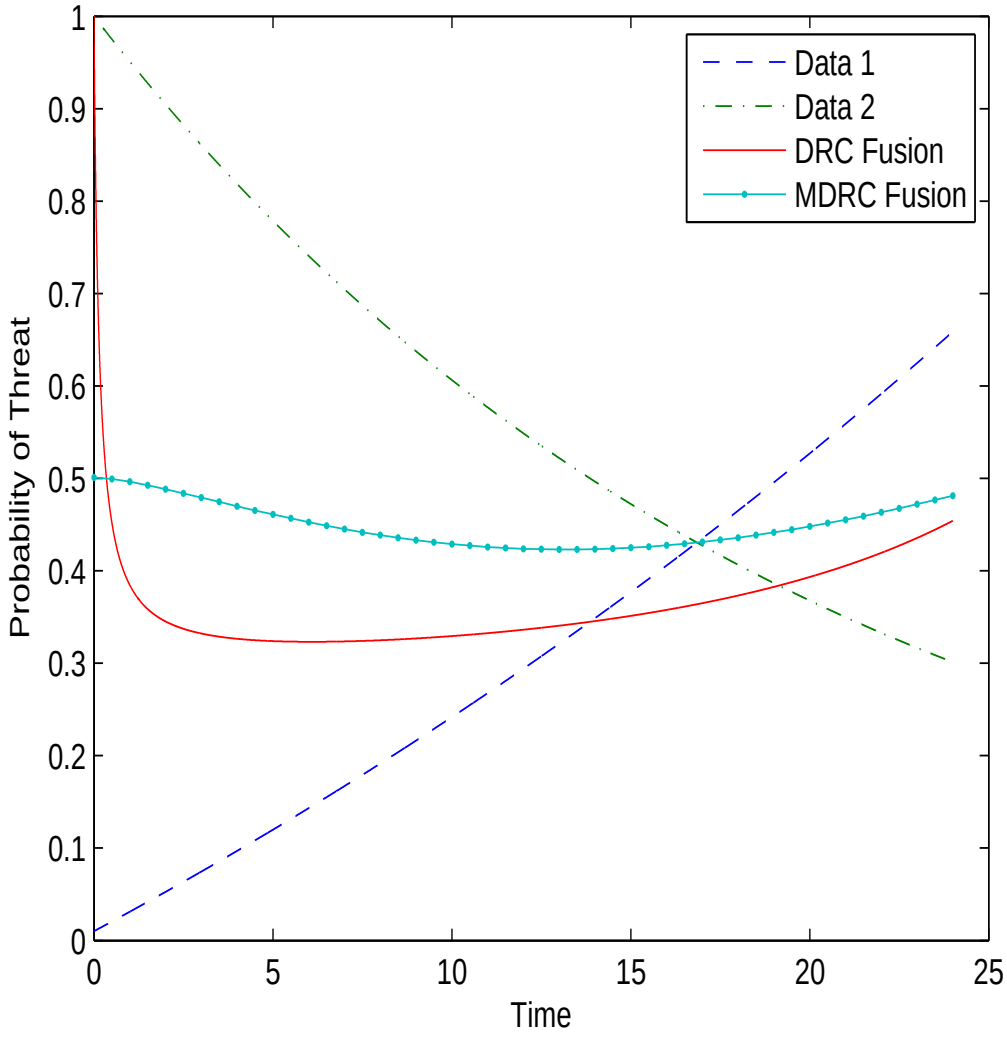

Figure 4.10: Data fusion scenario 2 - case 4

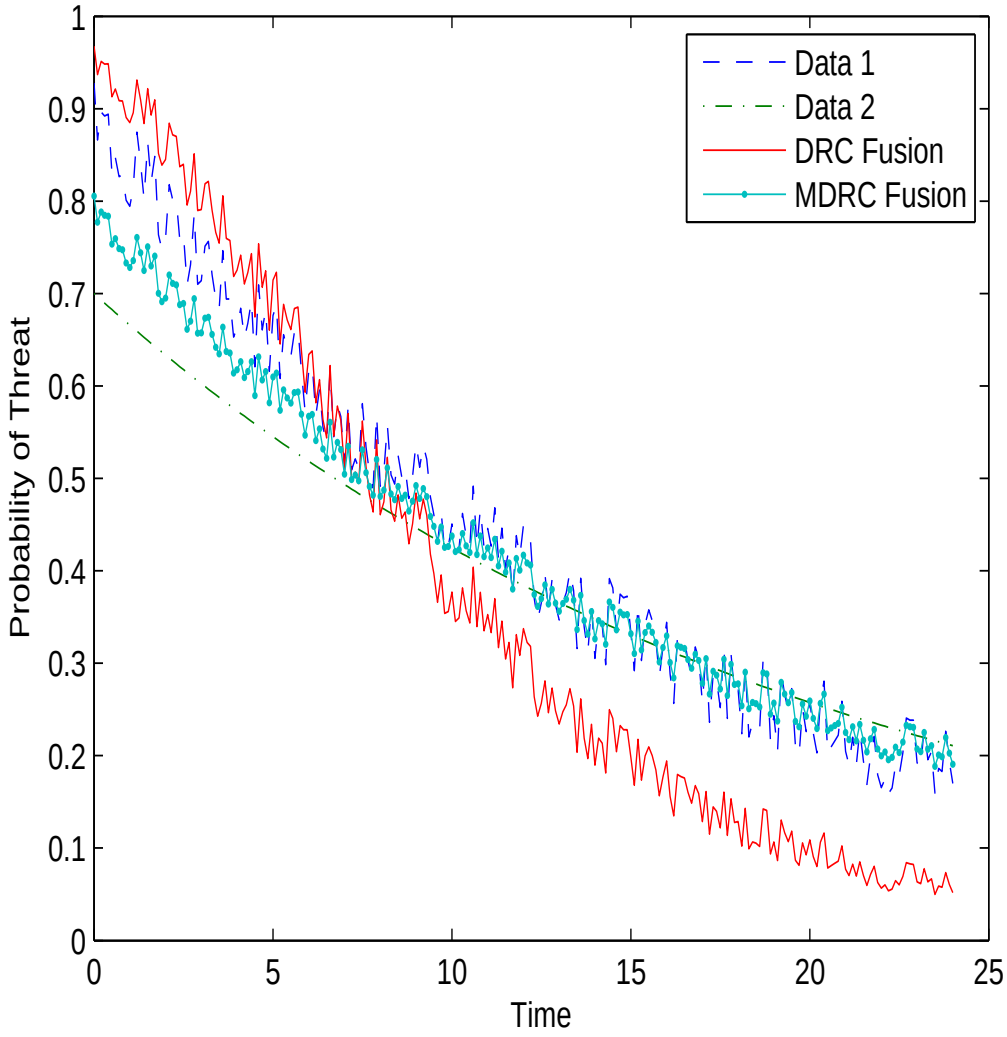

Figure 4.11: Data fusion scenario 2 - case 5

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In this paper, a novel approach for predicting events by fusing soft (e.g., Twitter) data with hard (e.g., AIS) data to improve the prediction capability was proposed. Various scenarios using the soft data to fuse with hard data is explained. The results for different types scenarios and its working are presented. This framework was demonstrated on a representative airborne surface surveillance environment however the proposed framework can be used for other surveillance applications as well. The novelty of the work was in the use of Modified Dempster's Rule of Combination to efficiently process large amounts of social network data along with large-scale AIS data and the preliminary results were presented. By using this Modified Dempster's rule of combination for the fusion process, all the conflicts between the data were removed. There is a plan to extend the work to consider more realistic scenarios with even larger data sets. In addition, theoretical performance quantification and computational complexity analysis are needed to be performed for assessing its efficiency.

5.2 Future Work

The proposed algorithms for fusing soft and hard data could be extended to data association problems where it is needed to process large amount of real time soft data and hard data. Instead of using the soft data as the probability values from the output of MDRC, can change it as state(X,Y), which can be easily incorporated with the hard data which is already in the state(X,Y). Another plan to take the proposed algorithm to next level is to implement it with other types of hard data other than AIS data along with the available social network data. This helps to check the efficiency of proposed algorithm on various other hard and soft data systems. It will take the proposed algorithm for fusion to next level.

Bibliography

[1] Sayandeep Acharya and Moshe Kam. Evidence combination for hard and soft sensor data fusion. In *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, pages 1–8. IEEE, 2011.

[2] Stephan Baas, Selvaraju Ramasamy, Jenny Dey DePryck, and Federica Battista. *Disaster risk management systems analysis: A guide book*, volume 3. Food and Agriculture Organization of the United Nations, 2008.

[3] Yaakov Bar-Shalom, Peter K Willett, and Xin Tian. Tracking and data fusion. *A Handbook of Algorithms. Yaakov Bar-Shalom*, 2011.

[4] Jeffrey A Barnett. Computational methods for a mathematical theory of evidence. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pages 197–216. Springer, 2008.

[5] Neil Bomberger, Bradley J Rhodes, Michael Seibert, Allen M Waxman, et al. Associative learning of vessel motion patterns for maritime situation awareness. In *Information Fusion, 2006 9th International Conference on*, pages 1–8. IEEE, 2006.

[6] Steven C Boraz. Maritime domain awareness: Myths and realities. Technical report, DTIC Document, 2009.

[7] Hermann Borotschnig, Lucas Paletta, and Axel Pinz. A comparison of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition. *Springer Computing*, pp. 293–319, 1999.

[8] Marc Cheong and Vincent CS Lee. A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter. *Information Systems Frontiers*, 13(1):45–59, 2011.

[9] Daniel Danu, Abhijit Sinha, Thiagalingam Kirubarajan, Mohammad Farooq, and Dan Brookes. Fusion of over-the-horizon radar and automatic identification systems for overall maritime picture. pages 1–8, 2007.

[10] Javier Diaz, Maria Rifqi, and Bernadette Bouchon-Meunier. A similarity measure between basic belief assignments. pages 1–6, 2006.

[11] Yi-jie Ding, She-liang Wang, Zhao Xin-dong, Lv Miao, and Yang Xi-qin. A modified dempster-shafer combination rule based on evidence ullage. 3:387–391, 2009.

[12] Zhu Feixiang. Mining ship spatial trajectory patterns from ais database for maritime surveillance. In *Emergency Management and Management Sciences (ICEMMS), 2011 2nd IEEE International Conference on*, pages 772–775. IEEE, 2011.

[13] Dale Fixsen and Ronald PS Mahler. The modified dempster-shafer approach

to classification. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 27(1):96–104, 1997.

[14] David L Hall, James Llinas, and Michael McNeese. Modeling and mapping of human source data. 2011.

[15] David Lee Hall and Sonya AH McMullen. *Mathematical techniques in multisensor data fusion*. Artech House, 2004.

[16] Deqiang Han, Chongzhao Han, and Yi Yang. A modified evidence combination approach based on ambiguity measure. In *Information Fusion, 2008 11th International Conference on*, pages 1–6. IEEE, 2008.

[17] Abbas Harati-Mokhtari, Alan Wall, Philip Brooks, and Jin Wang. Automatic identification system (ais): data reliability and human error implications. *Journal of navigation*, 60(03):373–389, 2007.

[18] Weiming Hu, Xuejuan Xiao, Zhouyu Fu, Dan Xie, Tieniu Tan, and Steve Maybank. A system for learning statistical motion patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(9):1450–1464, 2006.

[19] Renato Iannella and Karen Henricksen. Managing information in the disaster coordination centre: Lessons and opportunities. In *Proceedings of the 4th International ISCRAM Conference (B. Van de Walle, P. Burghardt and C. Nieuwenhuis, eds.)*, pages 1–11, 2007.

[20] Akshaya Iyengar, Tim Finin, and Anupam Joshi. Content-based prediction of temporal boundaries for events in twitter. In *Privacy, Security, Risk and Trust*

(PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, pages 186–191. IEEE, 2011.

[21] Anne-Laure Jousselme, Chunsheng Liu, Dominic Grenier, and Éloi Bossé. Measuring ambiguity in the evidence theory. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 36(5):890–903, 2006.

[22] Bahador Khaleghi, Alaa Khamis, and Fakhreddin Karray. Random finite set theoretic based soft/hard data fusion with application for target tracking. pages 50–55, 2010.

[23] Thiagalingam Kirubarajan, Yaakov Bar-Shalom, Krishna R Pattipati, and Ivan Kadar. Ground target tracking with variable structure imm estimator. *Aerospace and Electronic Systems, IEEE Transactions on*, 36(1):26–46, 2000.

[24] Krishnan Krishanth, Ratnasingham Tharmarasa, Thiagalingam Kirubarajan, Pierre Valin, and Eric Meger. Prediction and retrodiction algorithms for path-constrained targets. *Aerospace and Electronic Systems, IEEE Transactions on*, 50(4):2746–2761, 2014.

[25] Rudolf Kruse, Erhard Schwecke, and Jochen Heinsohn. *Uncertainty and vagueness in knowledge based systems: numerical methods*. Springer Science & Business Media, 2012.

[26] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, et al. The arab spring— the revolutions were tweeted: Information flows during the

2011 tunisian and egyptian revolutions. *International journal of communication*, 5:31, 2011.

[27] Ronald PS Mahler. Combining ambiguous evidence with respect to ambiguous a priori knowledge. i. boolean logic. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 26(1):27–41, 1996.

[28] Ronald PS Mahler. *Statistical multisource-multitarget information fusion*. Artech House, Inc., 2007.

[29] Ronald PS Mahler. *Advances in statistical multisource-multitarget information fusion*. Artech House, 2014.

[30] Pratap Misra and Per Enge. *Global Positioning System: Signals, Measurements and Performance Second Edition*. Lincoln, MA: Ganga-Jamuna Press, 2006.

[31] Benjamin Pannetier, Kaouthar Benameur, Vincent Nimier, and Michele Rombaut. Ground moving target tracking with road constraint. In *Defense and Security*, pages 138–149. International Society for Optics and Photonics, 2004.

[32] Daniela Pohl, Abdelhamid Bouchachia, and Hermann Hellwagner. Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 683–686. ACM, 2012.

[33] Kamal Premaratne, Manohar N Murthi, Jinsong Zhang, Matthias Scheutz, and Peter H Bauer. A dempster-shafer theoretic conditional approach to evidence updating for fusion of hard and soft data. pages 2122–2129, 2009.

[34] Angel Rabasa. *The lessons of Mumbai*, volume 249. Rand Corporation, 2009.

[35] Bradley J Rhodes, Neil Bomberger, Majid Zandipour, et al. Probabilistic associative learning of vessel motion patterns at multiple spatial scales for maritime situation awareness. In *Information Fusion, 2007 10th International Conference on*, pages 1–8. IEEE, 2007.

[36] Branko Ristic, B La Scala, Mark Morelande, and Neil Gordon. Statistical analysis of motion patterns in ais data: Anomaly detection and motion prediction. In *Information Fusion, 2008 11th International Conference on*, pages 1–7. IEEE, 2008.

[37] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[38] Augusto Dias Pereira dos Santos, Leandro Krug Wives, and Luis Otavio Alvares. Location-based events detection on micro-blogs. *arXiv preprint arXiv:1210.4008*, 2012.

[39] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[40] Kurt D Schwehr, Philip McGillivary, et al. *Marine Ship Automatic Identification System (AIS) for enhanced coastal security capabilities: an oil spill tracking application*. IEEE, 2007.

[41] Glenn Shafer et al. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.

[42] Elisa Shahbazian, Pierre Bergeron, Jean-Remi Duquet, Alexandre Jouan, and Pierre Valin. Data fusion applications for military and civilian purposes developed on dnd/lm canada decision support testbed. 1:420–424, 1999.

[43] Peter J Shea, Tim Zadra, Dale M Klamer, Ellen Frangione, and Rebecca Brouillard. Improved state estimation through use of roads in ground tracking. In *AeroSense 2000*, pages 321–332. International Society for Optics and Photonics, 2000.

[44] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[45] Philippe Smets and Robert Kennes. The transferable belief model. *Artificial intelligence*, 66(2):191–234, 1994.

[46] A Srikanth. Social media can solve many problems during natural disasters.

[47] Commander Brian J Tetreault. Use of the automatic identification system (ais) for maritime domain awareness (mda). In *OCEANS, 2005. Proceedings of MTS/IEEE*, pages 1590–1594. IEEE, 2005.

[48] Min Han Tun, Graeme S Chambers, Tele Tan, and Thanh Ly. Maritime port intelligence using ais data. *Recent advances in security technology*, page 33, 2007.

[49] Edward Waltz, James Llinas, et al. *Multisensor data fusion*, volume 685. Artech house Norwood, MA, 1990.

[50] Sarita Yardi and Danah Boyd. Dynamic debates: An analysis of group polarization over time on twitter. *Bulletin of Science, Technology & Society*, 30(5):316–327, 2010.

[51] Surender Reddy Yerva, Hoyoung Jeung, and Karl Aberer. Cloud based social and sensor data fusion. pages 2494–2501, 2012.

[52] Hans-Jürgen Zimmermann. *Fuzzy set theoryand its applications*. Springer Science & Business Media, 2001.