Improving the Accuracy of a Sequential Mining Model for Hurricane Trajectory Prediction

Jasmin Bissonnette, Caden Marofke, Taylor Cox
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba
Email: {bissonnj, marofkec, coxt3}@myumanitoba.ca

***Abstract*—In the 21st century alone, hurricanes have caused billions of dollars worth of damage due to property destruction, personal danger, and civil unrest. The catastrophic impact of hurricanes and other tropical storms has motivated extensive research into hurricane forecasting. Hurricane trajectory prediction is a subfield of hurricane forecasting that traditionally uses meteorological analysis to determine the future path of a given hurricane. While meteorological methods of hurricane trajectory prediction have been effective in the past, little attention has been paid to hurricane trajectory prediction with respect to historical frequent pattern mining. In this work, we propose multiple dimensions for improving the accuracy of datamining-based hurricane trajectory prediction. We propose the use of Haversine distance, a real-world distance formula for improved hurricane pattern matching accuracy. We also propose the use of weighted data to favor recent hurricane trends when training the trajectory prediction model. Finally, we use training data exclusively from the years 1950 to 2000 to reduce the impact of historical inaccuracies in hurricane recording methodologies. After implementing all three improvement dimensions on a contemporary data mining model for hurricane trajectory prediction based on AprioriAll, results show a prediction-correctness rate 10.22% higher than the current state-of-the-art, at 75.22%.**

***Index Terms*—Hurricane Trajectory, AprioriAll, Prediction, Pattern Matching**

I. Introduction

Hurricanes are a specific type of tropical cyclone exclusive to the Atlantic basin [1]. Hurricanes are divided into five categories according the Saffir-Simpson scale, based on wind-speed and expected damage [2]. In the United States alone, hurricanes incur damages valued as high as $157BN [3]. The damages incurred by hurricanes do not only include property damage, but also personal danger and in some cases widespread civil unrest [4]. These damages motivate extensive research in the field of hurricane prediction analysis. Hurricane prediction analysis is traditionally divided into the subfields of hurricane intensity prediction and hurricane trajectory prediction. This paper focuses exclusively on the subfield of hurricane trajectory prediction. In particular, the goal of this paper is to investigate and improve on the existing-state-of-the art in hurricane trajectory prediction systems based on sequential datamining.

Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [5]. Data mining applies to an extremely wide range of applications, including clickstream mining, social network community detection, and artificial intelligence. The datamining subdiscipline of particular interest in this work is *sequential* data mining. Sequential data mining uses a database of chronologically ordered inputs to mine frequent patterns and interesting association rules. These association rules are used to make predictions on future data entries such as future customer transactions. More specifically, association rules generated by sequential data mining algorithms including AprioriAll, SPIRIT and PrefixSpan can be used in the field of hurricane trajectory prediction.

To the knowledge of the authors, the existing standard in hurricane trajectory prediction via data mining is developed by Dong et al in [6]. The work cited proposes a hurricane trajectory prediction model based on datamining (HTPDM) that uses a modified variant of AprioriAll, a common sequential mining algorithm [7]. HTPDM executes AprioriAll against historical hurricane trajectories from 1900 to 2000 to generate interesting trajec-

tory association rules. These association rules are compared against test data composed of hurricane trajectories from 2001 to 2008. HTPDM divides a given testing trajectory into two parts: the initial trajectory T_i and the terminal trajectory T_t. T_i is compared against all association rule antecedents until a sufficiently matching rule R is found based on a predetermined fitness function. The consequent of R is then compared against T_t using pattern-matching techniques to determine if the prediction was correct. The results of HTPM showed that the model was able to achieve a best-case correctness ratio of 65%. In this work, HTPDM will be developed and various strategies will be employed to improve the correctness ratio. Reducing the impact of historical factors, favoring the impact of recent trends, and improving the realistic nature of the hurricane trajectory prediction model result in a new model dubbed WARD-HTP (Weighted-Asset, Realistic-Distance Hurricane Trajectory Predictor). The strategies employed in WARD-HTP ultimately result in a correctness ratio of 75.22%, a 10.22% improvement over HTPDM.

The remainder of this paper is structured as follows: First, Section II gives a detailed survey of related work in the hurricane prediction and sequential mining disciplines. Next, section III will cover the general motivations of this work with respect to hurricane trajectory prediction, and the specific motivations of this work with respect to improving the current state-of-the-art. Section IV will provide a description of the problem space, which includes specific implementation considerations in hurricane trajectory prediction. Section V will give an exposition of the solutions proposed in WARD-HTP for increasing the correctness ratio in hurricane trajetory prediction. Following that, Section VI will detail the experimental setup, including the training data and other experimental variables. Section VII will evaluate the results of the experiments, which show that the ideal configurations of WARD-HTP result in a maximal correctness ratio of 75.22%. This work will then conclude with section VIII, capturing the conclusions and directions of future work.

II. Related Work

The development of an accurate hurricane trajectory prediction model is a problem which has received little attention in the data mining literature. Works including [8] and [9] have been completed to develop data-mining approaches to hurricane intensity prediction. These works however do not describe approaches to hurricane trajectory prediction. Hurricane trajectory prediction has been covered in previous meteorological and geographic information system (GIS) works. These types of hurricane trajectory forecasting models include those seen in [10] or [11]. The intersection between data mining and hurricane trajectory prediction has been sparsely investigated in the literature. Publications such as [12] proposed a clustering-based model for typhoon track prediction based on specific characteristics of historical data. Typhoons are then clustered into groups depending on the general trajectory patterns they exhibit. While the described work provided novel results in typhoon track clustering, it is not clear if the proposed solution is able to predict the trajectories of new storms. Additionally, the clusters developed apply specific to typhoons, which are waterborne storms native to the east-pacific seaboard. Such storm patterns may not be applicable to patterns found in hurricanes, which are waterborne storms of the west-atlantic seaboard. For this reason, the proposed clustering patterns were not used in this work.

The primary baseline of this work is [6], which uses a specific implementation of AprioriAll (first proposed in [7]) to determine frequent hurricane trajectory patterns. These trajectory patterns correspond to interesting trajectory association rules, which are used to train a prediction model. The implementation of AprioriAll used in [6] is also used as the foundational model for this work. AprioriAll transforms a transaction database into a *customer* database, where each entry contains a customer and an ordered sequence of purchased items. The ordered sequence database is then mined for frequent ordered patterns, or frequent sequences. The modification of Apriori proposed in [6] mines frequent ordered patterns and thus skips the transformation step of AprioriAll. The model proposed in this work uses historical hurricane trajectory

readings from HURDAT, the canonical hurricane database maintained by the National Hurricane Center [13]. The data in HURDAT2 is ordered into $\langle Hurricane, Trajectory \rangle$ pairs, omitting the need for the sequence mapping step required at initialization of AprioriAll.

III. Motivation

There are two categories of motivation behind this work. The first motivation of this work is to increase the total investigation completed in the hurricane trajectory prediction subdiscipline. Little research effort has been allocated to this particular problem space of high economic, and social impact. The authors expect this work to constitute a growing effort by academic and government researchers to improve existing methods of hurricane trajectory prediction in the interest of public safety and economic stability.

The second motivation behind this work is the specific goal of improving the correctness ratio in the current standard data mining model for hurricane trajectory prediction. As previously described, HTPDM had a best case trajectory prediction correctness ratio of 65%. This means that of all testing trajectory for a which a matching rule was successfully produced, 65% of such matches were correct predictions. When HTPDM accounts for testing trajectories for which no rule could be matched, the correctness ratio decreases by 7.5% to a final value of 57.5% correctness. This decrease in accuracy is expected at some hurricane trajectories are outliers which cannot be predicted by frequent pattern mining, as will be discussed further in this work.

The motivation of improving the correctness ratio in the model encompasses the challenges of improving the best-case correctness ratio (the correctness ratio when ignoring rule-miss incidents) and the worst-case correctness ratio (the correctness ratio when accounting for rule-miss incidents). To resolve the motivations of this work, a series of improvement strategies are proposed as applications to the current hurricane trajectory prediction model based on AprioriAll. In general, these strategies seek to improve the realism of the prediction model, reduce historical factors which may negatively affect the model's correctness reatio, and favor recent hurricane trajectory trends which are likely to be reflected in current and future hurricane patterns.

IV. Problem Description

In this section, the problem of achieving accurate hurricane trajectory prediction is explained in detail. This problem consists of multiple challenges which apply generally to sequential datamining, and specifically to the problem domain of hurricane location analysis. Many of the domain-specific challenges of hurricane trajectory prediction arise from the fact that hurricane points are tracked as a series of $\langle Latitude, Longtitude \rangle$ points, which do *not* directly correspond to traditional (x, y) coordinates, as latitude and longtitude are points imposed upon a sphere. Challenges including accurate distance evaluation are results of the transition from a Euclidean to a Polar coordinate system.

The exposition of the hurricane prediction problem domain is divided into three parts. The first aspect of the problem domain concerns the underlying sequential mining approach used. This includes a thorough description of AprioriAll, its intended usage, and the reaction of AprioriAll to its input parameters. Next, the subproblem of region discretization will be discussed. Since $\langle Latitude, Longtitude \rangle$ points constitute continuous space, hurricane trajectories are converted to sequences of discrete regions to enable the discovery of frequent patterns. Finally, problems involving hurricane trajectory matching will be discussed. These problems include pattern matching for determining best fit rules and pattern matching for determining trajectory prediction correctness.

A. AprioriAll

In HTPDM and WARD-HTP, the algorithm that generates trajectory prediction possibilities is based on AprioriAll. AprioriAll was first proposed at IBM research as a method for discovering frequent sequential patterns in transactional databases [7]. In its original implementation, AprioriAll is able to generate assocation rules describing sentences of the form "Customers who buy P may also buy Q". In the hurricane trajectory prediction problem, AprioriAll is used to generate association ruls describing sentences of the form "Hurricanes located at P may later be located at Q". In the context

of AprioriAll, hurricanes are considered $customers$ and their sequences of recorded points are considered $transactions$. Consider the high-level implementation steps of AprioriAll as follows:

1) Map the transaction database into a customer database.
2) Generate frequent sequences based on minimum support.
3) Generate association rules based on minimum confidence.

1) Mapping: In step (1) of AprioriAll, a given transaction database *D* is grouped by customers, and the frequent items purchased by customers in individual transcations are recorded in a temporary table. The contents of this temporary table are mapped to corresponding sequences to enable the mining of frequent transaction sequences from the original database. The intention of step (1) is to map *D* in its original $\langle TransactionID, CustomerID, Items\rangle$ database form into a database *D'* of the form $\langle CustomerID, ItemSequence\rangle$.

The hurricane trajectory model in this work does not implement step (1), as the training data from HURDAT2 was instead translated into a $\langle HurricaneID, CoordinateSequence\rangle$ database as a preprocessing step before any aspects of the model are executed. By directly translating the data instead of processing it with AprioriAll, all trajectories in the original database are preserved. This includes coordinates which may ultimately be infrequent. The sequential mapping of AprioriAll is performed as a preprocessing step so that the database used in the prediction model will always consist of hurricanes and their coordinate sequences. This allows the prepared (mapped) database to be mined directly without the need to complete the mapping step every time AprioriAll is executed.

Fig. 1. Sample of mapped Hurricane-Coordinates database

HurricaneID	LatLonSequence
AL081976	⟨ (26.0:-84.0),(25.3:-83.3),(24.7:-82.7) ⟩
AL091976	⟨ (31.7:-68.2),(33.4:-67.5),(35.2:-66.4) ⟩
AL101976	⟨ (14.0:-48.0),(14.0:-49.3),(14.0:-50.6) ⟩
AL111976	⟨ (12.5:-37.5),(13.0:-39.0),(13.5:-40.5) ⟩

The above table shows an example of how the trajectory database is structured after the mapping preprocessing step. Only two dimensions are needed in the database, being the hurricane ID and the sequence of latitude-longtitude coordinates. All hurricanes in HURDAT2 have an ID corresponding to the location (the Atlantic), the index ordered by appearance in the hurricane year, and the year itself. For example, hurricane Katrina of 2005 corresponds to AL122005. The preprocessing step proposed is executed exactly once, allowing the hurricane database to be mined without the initial grouping and mapping stage.

2) Pattern Generation:

3) Rule Generation:

B. Region Discretization

C. Trajectory Matching

V. Solution

A. Haversine Formula

B. Weighted Training Data

C. Training Data Year Interval

VI. Experimental Setup

VII. Experimental Results

VIII. Conclusions and Future Work

References

[1]
[2]
[3]
[4]
[5]
[6]
[7]
[8]
[9]
[10]
[11]
[12]
[13]
[14]
[15]