Implementation of Data Mining Techniques for Weather Report Guidance for Ships Using Global Positioning System

P.Hemalatha
M.E Computer Science And Engineering IFET College Of Engineering Villupuram

Abstract

This paper deals with the implementation of data mining methods for guiding the path of the ships. The implementation uses a Global Positioning System(GPS) which helps in identifying the area in which the ship is currently navigating. The weather report on that area is compared with the existing database and the decision is made in accordance with the output obtained from the Data Mining technique. This decision about the weather condition of the navigating path is then instructed to the ship. This paper highlights some statistical themes and lessons that are directly relevant to data mining and attempts to identify opportunities where close cooperation between the statistical and computational communities might reasonably provide synergy for further progress in data analysis**.**

GLOBAL POSITIONING SYSTEM(GPS) provides specially coded satellite signals that can be processed in a GPS receiver enabling the receiver to compute position, velocity and time. Satellites were first used in position finding in a simple but reliable 2D Navy system called 'Transit' which laid the ground work for a system-"The Global Positioning System" that is funded and controlled by US Dept of Defense (DOD).

1. INTRODUCTION:

DATA MINING:

Data Mining means decision-making and data extraction. It also generates prediction mechanism from the available history. This implementation uses the Classification Models of Data Mining techniques.Data mining is a process of inferring knowledge from such huge data. Data Mining has three major components

1. Clustering or Classification,
2. Association Rules and
3. Sequence Analysis.

In classification/clustering we analyze a set of data and generate a set of grouping rules which can be used to classify future data. An association rule is a rule which implies certain association relationships among a set of objects in a database. In this process we discover a set of association rules at multiple levels of abstraction from the relevant set(s) of data in a database.

In sequential Analysis, we seek to discover patterns that occur in sequence. This deals with data that appear in separate transactions (as opposed to data that appear in the same transaction in the case of association).

2. CLASSIFICATION MODEL:

In Data classification one develops a description or model for each class in a database, based on the features present in a set of class-labeled training data. There have been many data classification methods studied, including decision-tree methods, such as C4.5, statistical methods, neural networks, rough sets, database-oriented methods etc. Using the training set, the Classification attempts to generate the description of the classes and these descriptions help to classify the unknown records. In addition to the training set, we can also have a test data set which is used to determine the effectiveness of a classification. The goal of the Classification is to build a concise model called Decision Tree that can be used to predict the class of the records whose class label is not known.

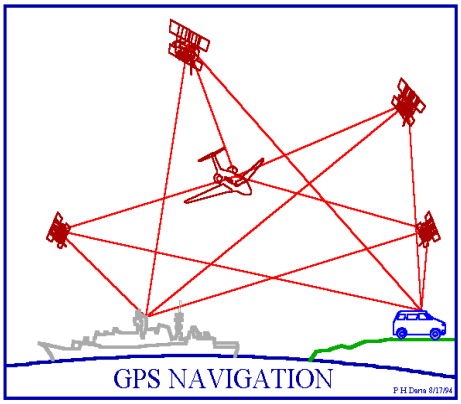
3. DECISION TREES:

A Decision tree is a Classification scheme, which generates a tree and a set of rules, representing the model of different classes, from a given data set. The set of records available for developing Classification methods is generally divided into two distinct subsets- a training set and a test set. The former is used for deriving the classifier, while the latter is used to measure the accuracy of the Classifier. The accuracy of the classifier is determined by the percentage of the test examples that are correctly classified. This implementation

uses the ID3 algorithm of the Classification Model to generate the Decision Tree.

4. The Global Positioning Systems

The US Global Positioning System (GPS) consists of a constellation of 24 satellites that emit radio signals for reception by specially designed devices. The GPS transmitter transmits the information regarding the latitude and longitude of the location where it is located to the satellite, which is then sent by the satellite to the receiver. If signals from one or more of these satellites are picked up by a GPS receiver, it can determine its location with high, reliable accuracy. The orbits are arranged such that there are, in fact, always at least 4 satellites visible from any point on the surface of the earth. Effectively, the satellite signal is continually marked with its own transmission time so that when received, the signal transit period can be measured with the synchronized receiver.



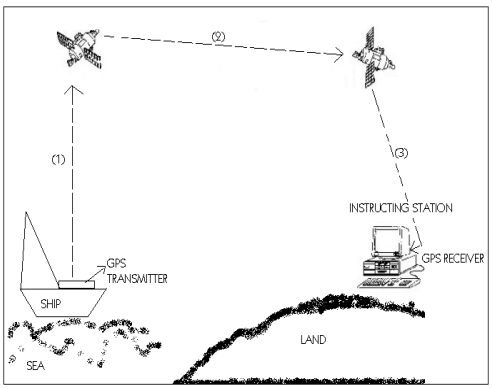
The satellites are so far out in space that the little distances we travel here on earth are insignificant. So if two receivers are fairly close to each other, say within a few hundred kilometers, the signals that reach both of them will have traveled through virtually the same slice of atmosphere, and so will have virtually the same.

5. Describing The Scenario

Steps involved:

- ❖The GPS transmitter is placed in the ship and its receiver is placed in the instructing station.
- ❖The analyzed weather report database is the training data.
- ❖The Decision Tree is constructed for the Training Data.
- ❖The GPS transmitter in the ship sends the information regarding the latitude and longitude of its current location to the nearby satellite.
- ❖This satellite in turn sends this information to the satellite which is closer to the instructing station.
- ❖The receiver present in the instructing station receives the GPS data from it.
- ❖The weather information for that particular location is collected.
- ❖The Decision Tree is traversed using this weather information and the required information is obtained.
- ❖This predicted decision is then sent to the ship and the ship navigates accordingly.

THE REAL PICTURE:



6. Construction Of Decision Tree

ID3 and C4.5 are algorithms introduced by Quinlan for inducing Classification Models, also called Decision Trees, from data. We are given a set of records. Each record has the same structure, consisting of a number of attribute/value pairs. One of these attributes represents the goal of the record, i.e. the attribute whose values are most significant to us. The problem is to determine a decision tree on the basis of answers to questions about the non-goal attributes predicts correctly the value of the goal attribute. Usually the goal attributes take only the values {true, false} or {success, failure}, or something equivalent. In any case, one of its values will mean failure.Here, we are dealing with the records reporting the weather conditions for instructing the ship for its safe navigation. The goal attribute specifies whether or not to move forward.

ATTRIBUTE	**POSSIBLE VALUES**
Climate	Sunny, Cloudy, Rainy
Temperature	Continuous
Humidity	Continuous
Stormy	True, False

The non-goal attributes are:

and the training data is:

Climate	**Temp (F)**	**Humidity(%)**	**Stormy?**	**Class**
Sunny	75	70	true	Safe
Sunny	80	90	true	Unsafe
Sunny	85	85	false	Unsafe
Sunny	72	95	false	Unsafe
Sunny	69	70	false	Safe
Cloudy	72	90	true	Safe
Cloudy	83	78	false	Safe
Cloudy	64	65	true	Safe
Cloudy	81	75	false	Safe
Rainy	71	80	true	Unsafe
Rainy	65	70	true	Unsafe
Rainy	75	80	false	Safe
Rainy	68	80	false	Safe
Rainy	70	96	false	Safe

Notice that in this example two of the attributes have continuous ranges, temperature and humidity. ID3 does not directly deal with such cases, though below we examine how it can be extended to do so. A decision tree is important not because we hope it will classify correctly new cases. Thus when building classification models one should have both training data to build the model and test data to verify how well it actually works.

7. Basic Ideas Behind Id3:

- In the decision tree each node corresponds to a non-goal attribute and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the goal attribute for the records described by the path from the root to that leaf. [This defines what is a decision tree].
- In the decision tree at each node should be associated the non-goal attribute which is most informative among the attributes not yet considered in the path from the root. [This establishes what a good decision tree is].
- Entropy is used to measure how informative is a node. [This defines what we mean by "good"].

7.1. Definitions:

If there are n equally probable possible messages, then the probability p of each is 1/n and the information conveyed by a message is –log (p) = log (n). [All logarithms are in base 2]. That is, if there are 16 messages then log (16)=4 and we need 4 bits to identify each message.

In general, if we are given a probability distribution P = (p1, p2…pn) the information conveyed by this distribution, also called the Entropy of P, is: I(P) = - (p1 * log(p1) + p2 * log(p2) +….+ pn * log(pn))

For example, if P is (0.5,0.5) then I(P) is 1, if P is (0.67,0.33) then I(P) is 0.92, if P is (1,0) then I(P) is 0. The more uniform is the probability distribution, the greater is its information.

If a set T of records is partitioned into disjoint exhaustive classes C1,C2,……,Ck on the basis of the value of the goal attribute, then the information needed to identify the class of an element of T is info(T) = I(P), where P is the probability distribution of the partition (C1,C2,…,Ck):

P = (|C1| / |T|, |C2| / |T|,……, |Ck| / |T|)

In the training set T, Info (T)=I (9/14,5/14)=0.94.

If we first partition T on the basis of the value of a non-goal attribute X into sets T1, T2 ,…,Tn then the information needed to identify the class of an element of T becomes the weighted average of the information needed to identify the class of an element of Ti, i.e., the weighted average of Info(Ti) : n

Info(X,T)= □ {(|Ti| / |T|) * Info(Ti)}

Here,

Info (Climate ,T) = 5/14 * I(2/5,3/5) + 4/14 * I(4/4,0) + 5/14 * I(3/5,2/5) = 0.694

Consider the quantity Gain(X,T) defined as Gain(X,T)=Info(T) – Info(X,T)

This represents the difference between the information needed to identify an element of T and the information needed to identify an element of T after the value of attribute X has been obtained, that is, the gain in information due to attribute X.

In our example, for the Climate attribute the Gain is

Gain(Climate, T) = Info(T) – Info(Climate, T) = 0.94 – 0.694 = 0.246. Similarly,

Gain (Humidity, T) = 0.151

Gain(Stormy, T) = 0.048

Gain(Temperature, T) = 0.029

We can use this notion of gain to rank attributes and to build decision trees where at each node is located the attribute with greatest gain among the attributes not yet considered in the path from the root.

The intent of this ordering is to create small decision trees so that the records can be identified after only a few questions.

7.2 The Id3 Algorithm:

The ID3 algorithm is used to build a decision tree, given a set of non-goal attributes C1,C2,…,Cn, the goal attribute C, and training set T of records. functionID3(R: a set of non-goal attributes,

C: the goal attribute,

S: a training set) returns a decision tree; Begin If S is empty, return a single node with value Failure;

If(S consists of records all with the same value for the goal attribute),

return a

single node with that value;

If R is empty, return a single node with as value the most frequent of the values of the goal attribute that are found in records of S; [note that there will be errors, that is, records that will be improperly classified];

Let D be the attribute with largest Gain(D,S) among attributes in R;
Let {dj | j=1,2,…,m} be the values of attribute D;

Let {Sj | j=1,2,…,m} be the subsets of S consisting respectively of records with value dj for attribute D;

Return a tree with root labeled D and arcs labeled d1,d2,…,dm going respectively to the trees
ID3(R - {D}, C, S1), ID3(R - {D}, C, S2),…, ID3(R - {D}, C, Sm);

End ID3;

8. Conclusion:

The weather report of the ship's location is made to traverse the Decision Tree and the corresponding decision is passed to the ship for its safe navigation. Thus this implementation, which uses many advanced concepts such as Data Mining and Global Positioning Systems, can also be extended for Aircrafts, Vehicle Tracking, Submarines, etc.

References:

[1] UCLA Data Mining Laboratory
[2] http://nugget.cs.ucla.edu:8001/main.html
[3] IBM QUEST Data Mining Project http://www.almaden.ibm.com/cs/quest/index.html
[4] Data Mining at Dun & Bradstreet
[5] http://www.santafe.edu/~kurt/text/wp9501/wp9501.shtml
[6] GPS Overview, by Peter.H.Dana. http://www.colorado.edu/geography/gcraft/notes/gps/gps_f.html

DECISION TREE FOR THE ABOVE ILLUSTRATION: