**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**
**IMPACT FACTOR: 6.017**

*IJCSMC, Vol. 6, Issue. 1, January 2017, pg.84 – 91*

# BIG DATA ANALYSIS USING SVM AND K-NN DATA MINING TECHNIQUES

**Dr. B. Lavanya*, B. Divya**

*Corresponding Author: lavanmu@gmail.com

Department of Computer Science, University of madras, Chennai, India

***ABSTRACT: Abstracting useful information from a big data has always been a challenging task. Data mining is a powerful technology with great potential to extract knowledge based information from such data. Prediction can be done with past and related records in different fields. Risk and safety have always been an important consideration in the field of aircraft. Prediction of accident in aircraft will save life and cost. This paper proposes an accident prediction system with huge collection of past records by applying effective predictive data mining techniques like Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN) which have a greater capacity to handle huge and noisy data that are used to predict accidents with more accuracy. The methods used, prove to handle noisy, unrelated and missing data. The prediction results are tabulated and ranges between 85% to 90%.***

***Keywords: Big data, SVM, K-NN, Accident Prediction.***

## 1. INTRODUCTION

This work focus on using data mining techniques in the process of accident prediction with aircraft accident details as training data set. Data from National Transportation Safety Board (NTSB), which records all the aircraft accidents, is used as training dataset for the proposed system. Various attributes which caused the accidents are analyzed. Collectively a set of ten attributes with one year of accident records are used as the training set. Huge Collection of past accident records of all types are available in different formats and found to be erroneous. Fields and records which support the aircraft accident prediction are filtered.

On the normalized data, SVM is applied to predict the future possibility of accident occurrence. SVM implements mapping of inputs into a high dimensional space using a set of nonlinear basis functions. For cross validation or to measure the accuracy level of SVM, two further techniques are implemented to predict accidents. K-NN which classifies real time data and versatile are also used in the prediction process.

This paper is organized as follows. Section 2 gives a detailed literature review, section 3 defines the problem, section 4 describes the dataset selection, section 5 draws the preprocessing stages, section 6 discuss the methods,

 

section 7 presents the accuracy, section 8 shows the implementation, section 9 draws the conclusion and finally section 10 presents future work.

## 2. LITERATURE REVIEW

Many researches have been done on the prediction of aircrafts accidents. We start the section by presenting some related works in prediction in general.

"Hybrid safety and analysis method based on SVM and RST", [1] *Ying Dai, Jin Tian, HaoRong, Tingdi Zhao*, research paper focused on providing a safe landing without an accident.

"Data mining approaches for aircraft accidents prediction", [2] by *A.B Arockia Christopher and Dr. S.Appavu*, used decision tree method to predict the warning level. Dataset used are pilot details, delay details, accident related details, maintenance details and flight details.

*A.B Arockia Christopher and Dr. S.Appavu*[3] analyzed various data preprocessing techniques to find best techniques which suits for airline data. Classification algorithms and many clustering techniques are used in for comparison. The data mining tool weka was used in this process. As a result of this analysis they have proved that Principal Components Attributes (PCA) Transformer would perform better than other attribute evaluators on airline data to reduce the dataset. On an empirical study on Turkey airline, decision tree technique is used to generate model. This model in turn is used to predict the warning level.

A similar research, "Analyzing Relationships between Aircraft Accidents and Incidents", [4] *ZohrehNazeri, George Donohue, Lance Sherry*, was done in USA. In this research various accidents and causes for these accidents are analyzed. Accident details from NTSB database and reasons for these accidents are maintained. Taxonomy was used for filtering the data. In order to maintain a uniform data structure data transformation is applied to transform the report into a vector containing common fields. Then STUCCO algorithm as used for finding the pattern. The result from the finding is then ranked using factor support ratio measure. Accuracy level of the output produced is also determined using accuracy algorithm.

Certain researches are done concentrating on particular attribute. For example weather. "The development of aircraft accident frequency model", [5] *D.K.Y Wong, D.E. Pitfield, R.E Caves and A.J Appleyard*, used weather as the only parameter. Temperature level, humidity, storm and wind speed are used as data set. Logistic regression analysis is used to estimate accident probability in a given weather condition.

In a research paper "*A system approach to accident causation in mining*", [6] Michael G. Lenne, Paul M. Salmon, Charles C. Liu, Margaret Trotter, analyzed human factors and classification systems (HFACS) were used to rise and caution level to any kind of accidents. Dataset from various accidents are stored in the database. Human error, technical faults, natural environment and climate conditions datasets were used.

## 3. PROBLEM DEFINITION

Understanding and processing an unstructured and dynamic data is a tedious work. Airline data are dynamic and seems to be unpredictable. Accidents and incidents may happen due to human or others natural calamities. Natural happenings such as bad weather, storm and birds happen suddenly which should be handled dynamically. Human errors can be avoided by proper planning and maintenance. Thus to process uncertain data, the usage of proper tools and techniques are needed.

## 4. DATASET SELECTION

Data from the database NTSB is used. Thousands of records from five to eight years are collected and stored for analysis. Pilot, runway, air speed and timing related data are collected and stored in the database.

The following attributes values are used as training data set, event start date, event end date, month, city, state, country, type (accident or incident or both), injury severity, aircraft details like category, make, model, registration, damage, number of engines, operation information like purpose of flight, schedule, air carrier and event details like airport name, code, weather condition, location of the accident, latitude, longitude.

 

## 5. DATA PREPROCESSING

Huge amount of data recorded from various accidents are available. National Transportation Safety Board (NTSB) [7] maintains a detailed record of all types of accident data. Field values may be missing or not available or not suitable for prediction or it may lead to wrong prediction. To overcome these problems data has to be pre-processed before starting the analysis. The process of cleaning involves filling of missing values, removing of unrelated attributes.

| | |
|---|---|
| $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ | (1) |

Data preprocessing is done on the available data. As the first step, data cleaning is done where the missing values are replaced by the attribute mean value.

Data filtering is done to extract recent records. Data sets for analysis contain hundreds of attributes, many of which are redundant and found to deviate the prediction process. Attribute subset selection reduces the data set size by pruning irrelevant and redundant attributes. At the end of data preprocessing stage a clean record is formed. The list of attributes considered in this work are weather condition, landing runway, pilot, injury severity, schedule, aircraft damage, aircraft category, engine type, location, and model as the key feature in accident prediction. Figure 1 depicts stage by stage data preprocessing steps.

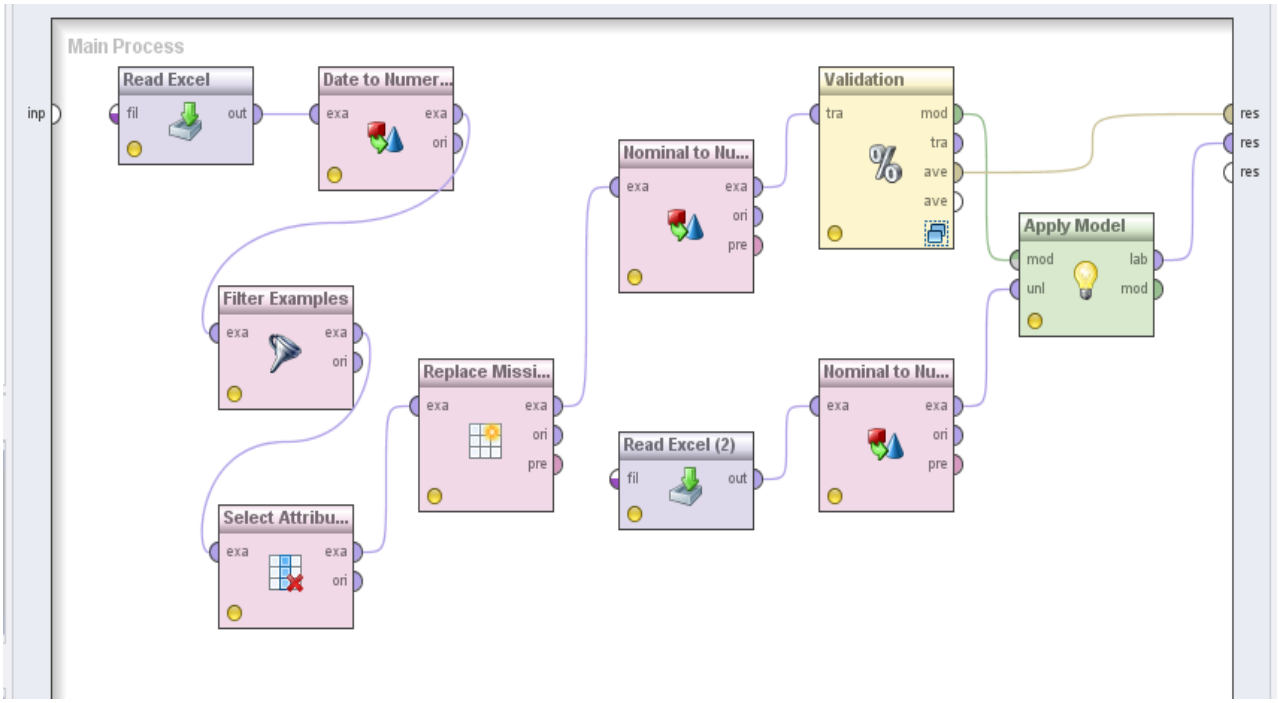


**Figure 1.** Data Preprocessing stages.

(a) Type conversion: Normalization, (b) Attribute selection: Removing unrelated or unwanted attributes, (c) Replace missing values: Removing or replacing the missing values with its mean value and (d) Data filtering: here we focus on the recent records and so we remove old records are the processes which a data has to undergo before applying any technique. Figure 1 explains the flow pictorially.

 

## 6. METHODS

Data mining has techniques to process unstructured and dynamic data. In this paper, prediction algorithms such as KNN (K-Nearest Neighbor) and SVM (Support Vector Machine) are used to predict the warning level.

### 6.1 SUPPORT VECTOR MACHINE

SVM is a promising method for the classification of both linear and non-linear data by a separating hyper-plane. It is a supervised learning technique from the field of machine learning applicable to both classification and regression. The appropriate set kernel function is chosen to handle high dimensional non-linear data set. The polynomial kernel function is used for classification. A multi-dimensional hyper-plane is formed which classifies the data as cause of accident and does not cause accident. The kernel function is given below.

$$K(x, y) = (x^{\mathsf{T}} y + c)^d \quad (2)$$

| Row No. | confidence(... | confidence(... | prediction(A... | WeatherCo... | WeatherCo... | WeatherCo... | LandingRu... | LandingRu... | LandingRu... |
|---|---|---|---|---|---|---|---|---|---|
| 49 | 0.617 | 0.383 | Yes | 0 | 1 | 0 | 1 | 0 | 0 |
| 50 | 0.578 | 0.422 | Yes | 0 | 1 | 0 | 1 | 0 | 0 |
| 51 | 0.801 | 0.199 | Yes | 0 | 1 | 0 | 1 | 0 | 0 |
| 52 | 0.775 | 0.225 | Yes | 0 | 1 | 0 | 1 | 0 | 0 |
| 53 | 0.555 | 0.445 | Yes | 0 | 1 | 0 | 0 | 1 | 0 |
| 54 | 0.515 | 0.485 | Yes | 0 | 1 | 0 | 0 | 1 | 0 |
| 55 | 0.757 | 0.243 | Yes | 0 | 1 | 0 | 0 | 1 | 0 |
| 56 | 0.727 | 0.273 | Yes | 0 | 1 | 0 | 0 | 1 | 0 |
| 57 | 0.408 | 0.592 | No | 0 | 1 | 0 | 0 | 0 | 1 |
| 58 | 0.370 | 0.630 | No | 0 | 1 | 0 | 0 | 0 | 1 |
| 59 | 0.633 | 0.367 | Yes | 0 | 1 | 0 | 0 | 0 | 1 |
| 60 | 0.595 | 0.405 | Yes | 0 | 1 | 0 | 0 | 0 | 1 |
| 61 | 0.617 | 0.383 | Yes | 0 | 1 | 0 | 1 | 0 | 0 |
| 62 | 0.578 | 0.422 | Yes | 0 | 1 | 0 | 1 | 0 | 0 |
| 63 | 0.801 | 0.199 | Yes | 0 | 1 | 0 | 1 | 0 | 0 |
| 64 | 0.775 | 0.225 | Yes | 0 | 1 | 0 | 1 | 0 | 0 |
| 65 | 0.555 | 0.445 | Yes | 0 | 1 | 0 | 0 | 1 | 0 |
| 66 | 0.515 | 0.485 | Yes | 0 | 1 | 0 | 0 | 1 | 0 |
| 67 | 0.757 | 0.243 | Yes | 0 | 1 | 0 | 0 | 1 | 0 |
| 68 | 0.727 | 0.273 | Yes | 0 | 1 | 0 | 0 | 1 | 0 |
| 69 | 0.408 | 0.592 | No | 0 | 1 | 0 | 0 | 0 | 1 |
| 70 | 0.370 | 0.630 | No | 0 | 1 | 0 | 0 | 0 | 1 |
| 71 | 0.633 | 0.367 | Yes | 0 | 1 | 0 | 0 | 0 | 1 |
| 72 | 0.595 | 0.405 | Yes | 0 | 1 | 0 | 0 | 0 | 1 |
| 73 | 0.694 | 0.306 | Yes | 0 | 0 | 1 | 1 | 0 | 0 |
| 74 | 0.659 | 0.341 | Yes | 0 | 0 | 1 | 1 | 0 | 0 |
| 75 | 0.850 | 0.150 | Yes | 0 | 0 | 1 | 1 | 0 | 0 |
| 76 | 0.829 | 0.171 | Yes | 0 | 0 | 1 | 1 | 0 | 0 |
| 77 | 0.637 | 0.363 | Yes | 0 | 0 | 1 | 0 | 1 | 0 |

**Figure 2.** Accident Prediction Using SVM.

Above Figure 2 shows the prediction result using SVM. The prediction (Yes/No) is determined based on the confidence level on No and Yes. Based on the attribute (weather, pilot, etc.,) values yes and no get confidence value from 0 to 1. Prediction is determined based on the confidence values as shown in Figure 2.

 

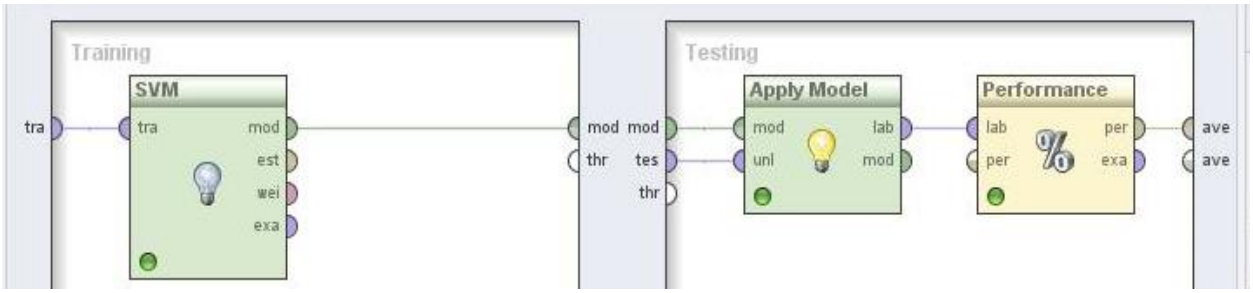


**Figure 3.**SVM and its Performance Measure Operators.

Figure 3 shows how SVM is used to predict the result from the test data. Each training set is applied as test set for better accuracy.

**TABLE 1.**Accuracy percentage of SVM after Dimensionality Reduction.

| | true Yes | true No | Class Precision |
|---|---|---|---|
| Pred. Yes | 74 | 1 | 98.67% |
| Pred. No | 1 | 32 | 96.97% |
| Class Recall | 98.67% | 96.97% | |

The predicted result by SVM under goes for the performance validation. The accuracy of result obtained is determined by using confusion matrix

### 6.1.2 DIMENSIONALITY REDUCTION

In dimensionality reduction, data encoding transformations are applied so as to obtain a reduced or compressed representation of the original data. Random Forest based approach, is applied for dimensionality reduction. It generates a large and carefully constructed set of trees against a target attribute and then uses each attribute's usage statistics to find the most informative subset of features. A statistical analysis is made on each tree and five major attributes which causes the accident are chosen for further prediction process. The attributes pruned are weather condition, landing runway, pilot, injury severity and schedule. With the selected attributes we are able to achieve greater accuracy (refer Table 1).

### 6.2 K-NEAREST NEIGHBOR (KNN)

K-nearest neighbor is an algorithm that stores all available cases and classifies new cases based on a similarity measure. KNN has been used in statistical estimation and pattern recognition as a non-parametric technique.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \qquad (3)$$

The above given Euclidean distance function is used for finding closeness of the attribute. Based on the closeness determined the classification of attributes to which class, it belongs to is determined. Value in K determines the

 

closeness measure. Here variable K holds the value 1 where the attributes are assigned to the class of its nearest neighbor.

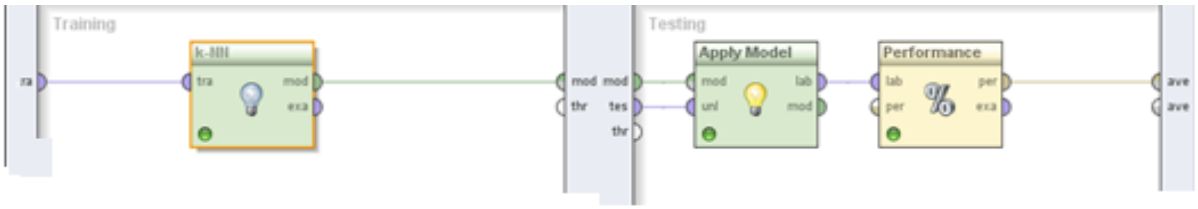


**Figure 4.** K- Nearest Neighbor and its Performance Measure Operators.

Figure 4 shows how the KNN operator used on test data to predict the result.

## 7. ACCURACY MEASURE

Accuracy is measured for each algorithm. Confusion matrix is used for calculating accuracy. It is a useful tool for analyzing the classifier to recognize the efficiency of the tuples of different classes. The error rate or misclassification rate of a classifier M, which is simply (1 - Accuracy) . Accuracy can be calculated as follows,

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FN + FP + TN)} \quad (4)$$

### 7.1 PERFORMANCE AND ACCURACY MEASURE

**Method 1 – Support Vector Machine**
**Accuracy: 96.27%**

**TABLE 2. Accuracy Percentage of SVM**

| | true Yes | true No | Class Precision |
|---|---|---|---|
| Pred. Yes | 74 | 3 | 96.10 % |
| Pred. No | 1 | 30 | 96.77 % |
| Class Recall | 98.67 % | 90.91 % | |

**Method 2 – K-Nearest Neighbor**
**Accuracy: 94.36 %**

 

**TABLE 3.Accuracy Percentage of K-NN**

| | true Yes | true No | Class Precision |
|---|---|---|---|
| Pred. Yes | 74 | 5 | 93.67 % |
| Pred. No | 1 | 28 | 96.55 % |
| Class Recall | 92.67 % | 84.85 % | |

### 7.2 ACCURACY COMPARISON OF SVM AND K-NN

**TABLE 5.Accuracy Measures of SVM and K-NN**

| | **True Yes** | **True No** |
|---|---|---|
| **SVM** | 98.67% | 90.91% |
| **K-NN** | 92.67% | 84.85% |

#### 7.2.1 COMPARSION CHART

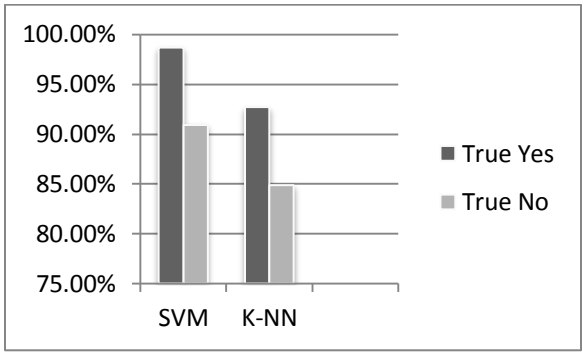


**Figure 5.** Accuracy comparsion chart of SVM and K-NN

Figure 5 depicts the accuracy comparison of algorithms implemented. The prediction level of yes and no by each alogithm is clear and SVM produced highest accuracy in prediction.

## 8.IMPLEMENTATION

Aircraft Accident Prediction is achieved through a powerful tool in data mining, called Rapid Miner. Rapid Miner is a software platform that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics.

Using Rapid Miner, SVM and K-NN have been implemented. Accuracy being a main part, it is estimated in each phase for better prediction.

## 9. CONCLUSION

Data mining efficiently acts here by predicting useful information from big, past and noisy data. Prediction of aircraft accident is a critical factor. Thus it has been implemented with three efficient data mining algorithms SVM and K-NN. These algorithms are efficiently used in medical sector to find cancer and kidney disease, crime prediction, weather prediction, banking sectors, marketing and in many other sectors. Three algorithms acts

 

efficiently in different real time conditions. On usage of these algorithms aircraft accident has been predicted with more accuracy.

### 10. FUTURE WORK

Data source from across the countries can be collected and analysed for better prediction. On successful implementation of aircraft accident prediction system, it can be further extended to railway and road ways. Implementation of these predictions system saves life and cost.

# REFERENCES

[1] Jin Tian, HaoRong, Tingd Zhao, "Hybrid Safety analysis method based on SVM and RST: An application to carrier landing of aircraft", **School of Reliability and Systems Engineering**, Vol. 80, Dec. 2015, Pages 56-65.
[2] A.B. Arockia Christopher, S. Appavu, "Data Mining Approaches for Aircraft Accidents Prediction", **Emerging Trends in Computing, Communication and Nanotechnology**, 2013, Pages 25-26.
[3] A.B. Arockia Christopher, S. Appavu, "Feature Selection for Prediction of Warning Level in Aircraft Accidents", **Advanced Computing and Communication Systems (ICACCS)**, 2013, Pages 1- 6.
[4] ZohrehNazeri, George Donohue, Lance Sherry, "Analyzing Relationships between Aircraft Accidents and Incidents", **International Conference on Research in Air Transportation (ICRAT)**, 2008 Pages 185-190.
[5] D.K.Y Wong, D.E Pitfield, R.E Caves and A.J Appleyard, "The Development of Aircraft Accident Frequency Models", **Safety and Reliability for Managing Risk – GuedesSoares**, 2006 Pages 83-90.
[6] Michael G.Lenne, Paul M. Salmon, Charles C. Liu, Margaret Trotter, "A System Approach to Accident Causation in Mining", **Accident Analysis and Prevention**. Vol. 48, Sep 2012, Pages 111-117.
[7] http://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx -For dataset reference.

 