Prediction of warning level in aircraft accidents using data mining techniques

A. B. Arockia Christopher
abchristomca@yahoo.com
Anna University
Chennai
India

S. Appavu alias Balamurugan
app_s@yahoo.com
KLN College of Information Technology
Sivagangai
India

ABSTRACT

Data mining is a data analysis process which is designed for large amounts of data. It proposes a methodology for evaluating risk and safety and describes the main issues of aircraft accidents. We have a huge amount of knowledge and data collection in aviation companies. This paper focuses on different feature selectwindion techniques applied to the datasets of airline databases to understand and clean the dataset. CFS subset evaluator, consistency subset evaluator, gain ratio feature evaluator, information gain attribute evaluator, OneR attribute evaluator, principal components attribute transformer, ReliefF attribute evaluatoboundar and symmetrical uncertainty attribute evaluator are used in this study in order to reduce the number of initial attributes. The classification algorithms, such as DT, KNN, SVM, NN and NB, are used to predict the warning level of the component as the class attribute. We have explored the use of different classification techniques on aviation components data. For this purpose Weka software tools are used. This study also proves that the principal components attribute with decision tree classifier would perform better than other attributes and techniques on airline data. Accuracy is also very highly improved. This work may be useful for an aviation company to make better predictions. Some safety recommendations are also addressed to airline companies.

Paper No. 4048. Manuscript received 1 August 2013, revised version received 17 March 2014, accepted 4 April 2014.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

1.0 INTRODUCTION

Data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships from large amounts of data stored in databases, data warehouses or other information repositories. Feature selection is an essential pre-processing step of data mining that helps increase the predictive performance of a model. The main aim of feature selection is to choose a subset of features with high predictive information and to eliminate irrelevant features with little or no predictive information. Classification is a data mining technique used to classify or predict number of individual collectively associates for data instances. The object of classification is to exactly predict the objective class for each case in the dataset. One of the commendable features of a classifier is its ability to tolerate noise. Its difficulty lies in handling quantitative data appropriately. Risk and safety have always been essential applications in aviation companies. Aviation accidents may result in human injury or even death. An airline company collects several case reports including structural and textual data. Institutions make important investments in order to collect such information from several sources.

In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features and irrelevant features provide no useful information in any context. Feature selection techniques provide three main benefits when constructing predictive models: improved model interpretability, shorter training times and enhanced generalisation by reducing over fitting. Feature selection is also useful as part of the data analysis process, as it shows which features are important for prediction and how these features are related. We have used different feature selection techniques for discovering attributes evaluator and generating a classification algorithm. These techniques are correlation feature selection subset evaluator, consistency subset evaluator, gain ratio feature evaluator, information gain attribute evaluator, OneR feature evaluator, principal components attribute transformer, ReliefF attribute evaluator and aymmetrical uncertainty attribute.

In this study, we applied different data mining methods on the incident reports. We use feature selection between attributes and classification algorithms to find the relations and rules about the incidents resulted in fatality. We also used Weka as a preprocessing tool to select the attributes which will be used in the analysis. This paper focuses on different feature selection techniques and applying traditional classification algorithms to understand and clean the dataset. Feature selection attributes are going to be used in this study in order to reduce the number of initial attributes. The classification techniques of data mining is used to predict the warning level of the component as the class attribute. All attributes reduced by find feature selections are more effective and give the rules that are found by these attributes. The new proposed model will be able to predict warning level of the accident for airline companies. We have explored the use of the different classification techniques on aviation components data. We offer the rules that were found by this analysis for the experts of aviation companies. Some safety recommendations are also addressed for airlines.

In this paper, five classification algorithms Decision Tree (DT), Naive Bayes Classification (NBC), Artificial Neural Network (ANN), *K*-Nearest Neighbour (KNN) And Support Vector Machines (SVM) have been considered for comparing their performance based on the aircraft data. In this experimentation, ten-fold cross-validation methods for comparison of selected classification algorithms have been used. This paper concentrates on performance of classification algorithms with different features combinations such as CFS, consistency, GR, information gain, OneR, PCA, ReliefF and symmetric with the selected datasets and also to improve the classification accuracy.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

The paper is organised as follows. In Sections 2 and 3 we describe the related work and selecting different features. In Section 4 we discuss the classification algorithms and in Section 5 we propose the proposed method. In Section 6 we describe the data sources and dataset descriptions and in Section 7 we produce the experimental results and discussion about classification with feature selection rule discovery for those aviation incidents which resulted in fatalities.

2.0 RELATED WORK

Data mining methods have been successfully introduced in many fields. It is still a research topic, but of tremendous interest to the industry in order to solve real-world problems[5]. Statistics, artificial intelligence techniques, decision tree approach, genetic algorithms and visualisation techniques are the major data mining techniques used to solve these problems. Risk and safety have always been important considerations in aviation. With the rapid growth in air travel, flight delays, cancellations and incidents/accidents have also dramatically increased in recent years[23]. Aviation accidents may result in human injury or even death. An airline company collects several case reports including structural and textual data. Institutions make important investments in order to collect such information from several sources. A database storing this information keeps essential data. But there exist some crucial problems in analysing the information. Nowadays, the analysis of such data cannot be done automatically and analysts have difficulties in dealing with the growing data efficiently and on time. In conclusion, in the automatic and smart analysis of the complex structure of high volume data in the aviation industry, capable instruments are needed. Data mining, an instrument which was not known before, could be potentially useful to reveal the information hidden in the data[8].

The aviation industry is one of these fields. With the rapid growth in air travel, flight delays, cancellations and incidents have also dramatically increased in recent years[23]. As a result of this, there is a large amount of knowledge and data accumulation in the aviation industry. This data could be stored in the form of pilot reports, maintenance reports, incident reports, component reports or delay reports. Also in the aviation industry, data mining applications has been performed. Bineid and Fielding[12] used data mining techniques to explain the development of a dispatch reliability prediction method for passenger aircraft. Nazeri and Zhang[23] described the application of data mining to analysing severe weather impacts on the national airspace system (CAA) was applied to accomplish the study. Their approach takes the more complex relationships among relevant performance.

The US Mine Safety and Health Administration (MSHA) developed a mine accident database from Part 50 of the Federal mine safety regulations. This database has been used to track the numbers, rates and severity of mine accidents in the United States. Epidemiologists and mine safety researchers still use it to perform many analyses, helping to guide research and best practice[16]. Dessureault *et al* explore the background of the Part 50 database, give a general background of data warehousing and data mining, and present some of the interesting analyses that resulted from a modernised Part 50 data warehouse using data mining.

Shyur[10] proposes a model that allows investigation of non linear effects of aviation safety factors and flexible assessment of aviation risk. The main goal of the research is to develop a model to provide relative risk probability inference and trend analysis among different kind of human errors which may cause any major aviation events. A subset of data gathered from the flight safety management information systems (FSMIS) developed by the office of the Taiwan Civil Aeronautics Administration takes aviation risk factors into account. Using the models presented, risk has been assessed as the probability of occurrence of a specific type of human error related aviation accident. The potential human error related risk could be identified and monitored timely.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

The results can provide better references to the civil aviation communities to manage the aviation safety risk, thus corrective action can be taken to reduce the occurrence of aviation accidents. It also allows finding fundamental cause of human error related accidents through the analysis of operational safety data. Finally, the results of the case study demonstrate that the proposed model is a more promising regression model with the potential of becoming very useful in practice.

Reference 18 also demonstrate how to use Data Mining (DM) analysis which can be applied to evaluate how well cameras that monitor red-light-signal controlled intersections to improve traffic safety by reducing fatalities. The paper denoted several different data modelling techniques — decision trees, neural networks, market-basket analysis and *K*-means models. Decision trees build rule sets that can affect future decision making. Neural networks try to predict future outcomes by looking at the effects of historical inputs. Market-basket analysis shows the strength of the relation-ships between variables. *K*-means models weigh the impact of homogenous clusters on target variables. All of these models are demonstrated using real data gathered by the US Department of Transportation from fatal accidents at red light signal-controlled intersections in Maryland and Washington DC from the year 2000 through 2003.

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Huge data sets have grown increasingly large in terms of number of dimensions and number of instances. Reducing the number of dimensions by selecting variables or features is effective in dealing with high-dimensional data. Variable and feature selection have became the focus of much research in areas of application for which data sets with tens or hundreds of thousands of variables are available. The objective of variable selection is threefold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data[6]. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalises better to unseen points.

A feature selection is usually meant as a process of finding a subset of features from the original set of features forming patterns in a given data set, optimal according to the defined goal and criterion of feature selection[19]. Many classification problems involve features whose specificity demand some form of feature space transformation (preprocessing) coupled with postprocessing consensus analysis in order to accomplish a successful discrimination between different classes. Pizzi and Pedrycz[15] present a new methodology, which systematically addresses these design classification issues in their study.

3.0. SELECTING DIFFERENT FEATURES

3.1 Feature selection

Feature selection is an important pre-processing method in data mining. Feature selection solves the scalability issue and increases the performance of classification models by eliminating redundant, irrelevant or noisy features from high dimensional datasets. Feature selection is a process of selecting a subset of relevant features by applying certain evaluation criteria. In general, the feature selection process consists of three phases. It starts with selecting a subset of original features and evaluating each feature's worth in the subset. Secondly, using this evaluation, some features in the subset may be eliminated or enumerated to the existing subset. Thirdly, it checks whether the final subset is good enough using certain evaluation criterion.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

Feature selection can be classified into feature subset selection and feature ranking. Feature ranking calculates the score of each attribute and then sorts them according to their scores. Feature subset selection selects a subset of attributes which collectively increases the performance of the model.

The number of features captured in the data is very large. We use the CFS subset evaluator, consistency subset evaluator, gain ratio feature evaluator, information gain attribute evaluator, oneR attribute evaluator, principal components attribute transformer, relief feature attribute evaluator and the symmetrical uncertainty attribute evaluator.

3.2 Feature selection techniques

In this work we focus on the following feature selection techniques:

1. Correlation feature selection
2. Consistency feature subset
3. Gain ratio
4. Information gain
5. OneR attribute
6. Principal component analysis
7. Relief feature
8. Symmetric uncertainty

3.2.1 Correlation feature selection

CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficients are used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation[5]. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search and genetic search.

The equation for CFS is given:

$$r_{zc} = \frac{k\overline{rz_i}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \quad \ldots (1)$$

where r_{zc} is the correlation between the summed feature subsets and the class variable, k is the number of subset features, r_{zi} is the average of the correlations between the subset features and the class variable, and r_{ii} is the average inter-correlation between subset features[2].

3.2.2 Consistency subset feature selection

The idea behind these measures is that, in order to predict the concept or class value of its instances, a data set with the selected features alone must be consistent[1]. That is, no two instances may have the same values on all predicting features if they have a different concept value. Therefore, the goal is equivalent to select those features that better allow defining consistent logical hypothesis about the training data set. As the higher the number of features, the more consistent hypothesis

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

that can be defined, the requisite, of a data set having consistency, is usually accompanied with the criterion of finding a small feature set. In any case, the search for small feature sets is the common goal of feature selection methods, so this is not a particularity of consistency based methods.

3.2.3 Gain ratio

The information gain measure is biased towards tests with many outcomes. That is, it prefers to select attributes having a large number of possible values over attributes with fewer values even though the later is more informative[2]. C4.5 a successor of ID3 uses an extension of information gain known as gain ratio, which attempts to overcome this bias.

Let Y be set consisting of y data samples with n distinct classes. The expected information needed to classify a given sample is given by:

$$I(Y) = -\sum_{i=1}^{n} p_i \log_2 (p_i) \quad \ldots (2)$$

where p_i is the probability that an arbitrary sample belongs to class C_i and is estimated by y_i/y.

Let attribute A has v distinct values. Let y_{ij} be number of samples of class C_i in a subset Y_j. Y_j contains those samples in Y that have value a_j of A. The entropy, or expected information based on the partitioning into subsets by A, is given by:

$$E(A) = -\sum_{i=1}^{n} I(Y) \frac{Y_{1i} + Y_{2i} + + Y_{ni}}{Y} \quad \ldots (3)$$

The encoding information that would be gained by branching on A is:

$$Gain(A) = I(Y) - E(A) \quad \ldots (4)$$

C4.5 uses gain ratio which applies normalisation to information gain using a value defined as:

$$SplitInfo_A(Y) = -\sum_{i=1}^{v} (|Y_i| / |Y|) \log_2 (|Y_i| / |Y|) \quad \ldots (5)$$

The above value represents the information generated by splitting the training data set S into v partitions corresponding to v outcomes of a test on the attribute A.

The gain ratio is defined as:

$$Gain\ Ratio(A) = Gain(A) / SplitInfo_A(Y) \quad \ldots (6)$$

The attribute with the highest gain ratio is selected as the splitting attribute.

3.2.4 Information gain

It is a commonly used measure in the fields of information theory and machine learning. IG measures the number of bits of information gained about the class prediction by knowing the value of a given feature when predicting the class[21].

The information gain of a given attribute X with respect to class attribute Y is the reduction in uncertainty about the value of Y when the value of X is known. The value of Y is measured by its entropy, $H(Y)$[14]. The uncertainty about Y, given the value of X is given by the conditional probability of Y given X, $H(Y|X)$.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

The entropy of Y (which consists of classes Y_1 and Y_2) is given by:

$$H(Y) = - \sum_{i=1}^{k} P(Y=Y_i) \log_2 (P(Y=Y_i)) \quad \ldots (7)$$

The conditional entropy of Y given X (consisting of values $X_1, X_2, \ldots, X_r$) is:

$$H(Y/X) = - \sum_{j=1}^{r} P(X = X_j) H(Y/X=X_j) \quad \ldots (8)$$

The information gain of feature X is defined as:

$$IG(\mathrm{X}) = H(Y) - \mathrm{H}(Y/X) \quad \ldots (9)$$

3.2.5 OneR attribute evaluation

Rule-based algorithms provide ways to generate compact, easy-to-interpret, and accurate rules by concentrating on a specific class at a time. One way of generating classification rules is to use decision trees. The disadvantage of using a decision tree is because it is complex and incomprehensible[17]. A classification rule can be defined as $r = (a, c)$ where a is a precondition which performs a series of tests that can be evaluated as true or false and c is a class that apply to instances covered by rule r.

A general rule of a rule-based algorithm tries to cover all instances belonging to a class. Rule base algorithms work on a specific class at a time. Rule-based algorithms follow three steps: Generate rule R on training data S, remove the training data covered by rule and repeat the process. OneR is the simplest approach to finding a classification rule as it generates one level decision tree. OneR constructs rules and tests a single attribute at a time and branch for every value of that attribute. For every branch, the class with the best classification is the one occurring most often in the training data.

3.2.6 Principal component analysis

Principal component analysis (PCA) is a standard technique used to handle linear dependence among variables. A PCA of a set of m variables generates m new variables (the principal components), $PC1 \ldots PCm$. Each component is obtained by linear combination of the original variables[14], that is:

$$PCi = \sum_{j=1}^{\mathrm{m}} bi,j.Xj \quad \ldots (10)$$

$$PC = B^T \overrightarrow{X} \quad \ldots (11)$$

Where Xj is the jth original variable, bi,j the linear factor. The coefficients for PCi are chosen so as to make its variance as large as possible. Mathematically, the variation of the original m variables is expressed by the covariance matrix. The transformation matrix **B**, containing the bi,j coefficients, corresponds to the covariance eigenvector matrix. Sorting the eigenvectors by their eigenvalues, the resulting principal components will be sorted by variance. In fact, the size of an eigenvalue defines how far a feature vector projected onto the eigenspace will be scaled along the correspondent eigenvector direction. Thus this new feature set is naturally ranked by variance which is useful if variance is a reasonable proxy for predictivness.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

3.2.7 Relief feature selection

It is efficient, aware of the contextual information, and can correctly estimate the quality of features in problems with strong dependencies between features. The key idea of the original ReliefF algorithm is to estimate the quality of features according to how well their values distinguish between instances that are near to each other.

Relief is an easy to use, fast and accurate algorithm even with dependent features and noisy data [9]. Relief works by measuring the ability of an attribute in separating similar instances. The process of ranking the features in relief follows three basic steps:

1. Calculate the nearest miss and nearest hit.
2. Calculate the weight of a feature.
3. Return a ranked list of features or the top k features according to a given threshold.

3.2.8 Symmetric uncertainty

Symmetric uncertainty is used to measure the degree of association between discrete features. It is derived from entropy [22]. It is a symmetric measure and can be used to measure feature-feature correlation.

$$SU = 2 \cdot 0 \text{ x } \frac{H(X)+H(Y)-H(X,Y)}{H(Y)+H(X)} \quad \ldots (12)$$

Symmetrical uncertainty is calculated by the above equation. $H(X)$ and $H(Y)$ represent the entropy of features X and Y. The value of symmetrical uncertainty ranges between 0 and 1. The value of 1 indicates that one variable (either X or Y) completely predicts the other variable[4]. The value of 0 indicates the both variables are completely independent.

4.0 CLASSIFICATION ALGORITHMS

Classification of data is very important task in data mining and machine learning. There are large number of classifiers that are used to classify the data such as bayes, function, rule based and tree, etc. The goal of classification is to correctly predict the value of a designated discrete class variable, given a vector of predictors or attributes.

The input to the problem is a data-set called the training set, which consists of a number of examples each having a number of attributes. The attributes are either continuous, when the attribute values are ordered, or categorical when the attribute values are unordered. One of the categorical attributes is called the class label or the classifying attribute. The objective is to use the training set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data-set. Classification has been studied extensively in statistics, machine learning, neural networks and expert systems over decades[20]. There are several classification methods:

- Decision tree algorithms
- Bayesian algorithms
- Rule based algorithms
- Neural networks

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

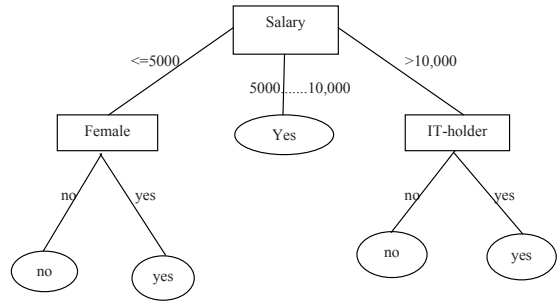


Figure 1. The typical decision tree model.

- Support vector machines
- Associative classification
- Distance based methods
- Genetic algorithms

After selecting the most different features, we apply DT, NB, SVM, KNN and NN to obtain error rates on our training samples. The classification results of these algorithms are then used to compare the effectiveness of various feature selection methods.

4.1 Decision trees

Decision tree induction[11] is a very popular and practical approach for pattern classification. Decision tree is constructed generally in a greedy, top down recursive manner. The tree can be constructed in a breadth first manner or depth first manner. Decision tree structure consists of a root node, internal nodes and leaf nodes. The classification rules are derived from the decision tree in the form of [if… then… else…]. These rules are used to classify the records with unknown value for class label. The decision tree is constructed in two phases: Building phase and Pruning phase. In the building phase of the tree the best attributes are selected based on attribute selection measures, such as information gain, gain ratio, Gini index, etc. Once the best attribute is selected then the tree is constructed with that node as the root node and the distinct values of the attribute are denoted as branches. The process of selecting best attribute and representing the distinct values as branches are repeated until all the instances in the training set belong to the same class label.

In the pruning phase the sub-trees are eliminated which may over fit the data. This enhances the accuracy of a classification tree. Decision trees handle continuous and discrete attributes. Decision trees are widely used because they provide human readable rules, easy to understand, construction of decision tree is fast and it yields better accuracy. There are several algorithms to classify the data using decision trees.

J48 are the improved versions of C4.5 algorithms or can be called as optimised implementation of the C4.5. The output of J48 is the decision tree. A decision tree is similar to the tree structure having root nodes, intermediate nodes and leaf nodes. Each node in the tree consists of a decision and that decision leads to our result. Decision trees divide the input space of a dataset into mutually exclusive areas, each area having a label, a value or an action to describe its data points. Splitting criterion is used to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node. Figure 1 shows the typical decision tree model using J48 for an aircraft dataset whether incident or not. Decision tree is formed by using the accident attribute of the aircraft dataset.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

4.2 Naive Bayes

A Naive Bayes (NB) classifier is a simple probabilistic classifier based on Bayes theorem where every feature is assumed to be class-conditionally independent[13]. In naïve bayes learning, each instance is described by a set of features and takes a class value from a predefined set of values. Classification of instances gets difficult when the dataset contains a large number of features and classes because it takes enormous numbers of observations to estimate the probabilities[13]. When a feature is assumed to be class-conditionally independent, it means that the effect of a variable value on a given class is independent of the values of other variables.

4.3 Support vector machines

SVMs are a kind of blend of linear modelling and instance-based learning. A SVM selects a small number of critical boundary samples from each class and builds a linear discriminant function that separates them as widely as possible. In the case that no linear separation is possible, the technique of 'kernel' will be used to automatically inject the training samples into a higher-dimensional space, and to learn a separator in that space.

A support vector machine (SVM) is a hyperplane that separates two different sets of samples with maximum distance of hyperplane to nearest samples from both sets [11]. The formula for the output of a linear SVM is

$$u = \overline{w}.\overline{\boldsymbol{x}} - b \qquad \ldots (13)$$

In this equation **w** is the normal vector to the hyperplane and **x** is the input vector. The nearest points lie on the planes $u = \pm 1$. The distance d is

$$d = \frac{1}{\|\mathbf{w}\|^2} \qquad \ldots (14)$$

The maximum distance d can be expressed using optimisation problem

$$min_{\overline{w}}, b \ \tfrac{1}{2} \ \|\mathbf{w}\|^2 \ \text{subject to } yi \ (\overline{\mathbf{w}}.\overline{\mathbf{x}}_i - b) \geq 1 \qquad \ldots (15)$$

where x_i is the ith training sample and y_i is the correct output of the SVM for the ith training sample. The value y_i is +1 for the positive samples and -1 for the negative samples.

4.4 *K* – Nearest neighbour

KNN is one of the supervised learning algorithms that have been used in many applications in the field of data mining, statistical pattern recognition and many others. It follows a method for classifying objects based on closest training examples in the feature space. An object is classified by a majority of its neighbours. K is always a positive integer. The neighbours are selected from a set of objects for which the correct classification is known.

The KNN classifier is a non parametric lazy learning algorithm. A data sample in KNN is classified on the basis of a selected number of k nearest neighbour[7]. The assumptions followed in KNN are:

1. KNN assumes that the data is in a feature space, so they have the concept of distance. Euclidean distance can be used to compute distance between vectors.
2. Each training vector is associated with set of vectors and class label.
3. K decides how many neighbours influence the classification

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

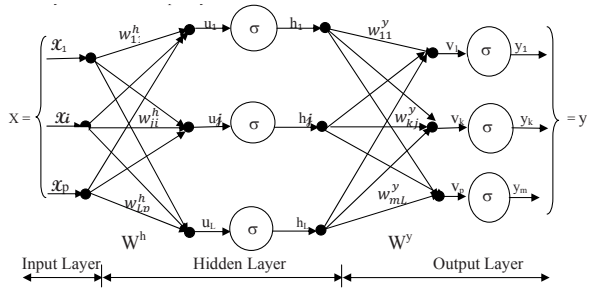


Figure 2. The perceptron network with three layers.

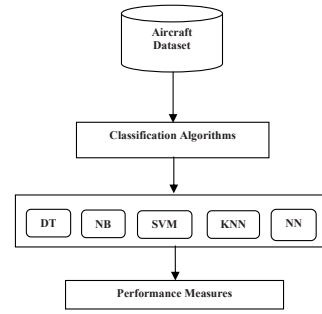


Figure 3. Block diagram of the classification model.

The following rule is the majority rule that is used extensively in KNN. The classification of the nearest neighbours can be decided by calculating the count of individual class values from all k nearest neighbours. The class value with the majority count is classified to the sample. K is an odd number to avoid duplicate counts.

4.5 Multi layer perceptron

Multi layer perceptron can be defined as neural network and artificial intelligence without qualification. A multi layer perceptron (MLP) is a feed forward neural network with one or more layers between input and output layer. The following diagram (Fig. 2) illustrates a perceptron network with three layers: Each neuron in each layer is connected to every neuron in the adjacent layers. The training or testing vectors are presented to the input layer, and processed by the hidden and output layers.

5.0 PROPOSED METHOD

Classification of data is a very important task in data mining. We propose to improve classification accuracy by different feature selection techniques and different classification methods. Feature selection helps to identify the attributes that are most important in predicting a particular outcome. Feature selection techniques are used for selecting subset of relevant features from the dataset to build robust learning models. Figure 3 shows the block diagram of the classification model before applying the feature selection. Figure 4 is the block diagram of the new proposed model illustrating the data mining methods after applying the features selection methods in to the aircraft dataset analysis problem. This model is proposed for improving classification accuracy by combining the prediction of multiple classifiers. The different data mining classifiers such as DT, NB, SVM, KNN and NN are used for classification. The aircraft dataset investigated in this model is taken from different aviation companies which are classified under two categories Safety and Risk. This model has ten-fold cross validation as a test method. The classification models are evaluated using the different performance metric. A comparative study is carried on the performances of the different classifiers and after carrying out different feature selection techniques the classifiers accuracy significantly improved. The performance of each features selection methods and classifiers model is evaluated by using statistical measures like accuracy, specificity and sensitivity. The receiver operating characteristics (ROC) chart is also used for measuring performances of the whole dataset.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

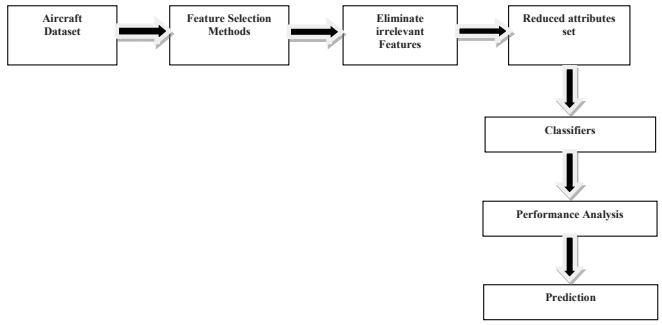


Figure 4. Block diagram of the proposed model.

5.1 Feature selection

We proposed a different feature selection method in this model. The number of features obtained in the dataset is very large. We are applied the number of different techniques CFS subset evaluator, consistency subset evaluator, gain ratio feature evaluator, Information gain attribute evaluator, OneR attribute evaluator, principal components attribute transformer, ReliefF attribute evaluator and symmetrical uncertainty attribute evaluator. In this model found that the feature selection methods attribute reduces more number of redundant and irrelevant attributes thereby increases the performance of classification methods.

5.2 Classification

We applied five classification algorithms for aircraft dataset such as DT, NB, SVM, KNN and NN. First we applied these algorithms for entire dataset. The aircraft ARFF will contain large quantity of data and applying classification algorithms to this dataset is time consuming and also gives result with less accuracy. Hence we have to reduce the number of initial attributes by using different feature attribute selection methods. After the elimination and reduce the attribute is fed into the five classifications algorithm and which algorithm is best accuracy for this prediction is investigated. Likewise, all other attribute selection and classification algorithms are applying for aircraft dataset. From that we concluded that the performance of classification accuracy after applying feature selection attribute methods is significantly improved in this model.

5.3 ROC chart

A ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Here, to be improving specificity and sensitivity in our proposed model.

Figure 4 illustrates the data mining methods applied to the aircraft data analysis problem. The following steps are involved

Step 1: Aircraft new dataset is given as input.

Step 2: Apply feature selection methods in initial data.

Step 3: Eliminate irrelevant features attributes

Step 4: Reduced attributes set after eliminate.

Step 5: Evaluate the classifiers performance.

Step 6: Performance of analysing the data

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

6.0 DATA SOURCES AND DESCRIPTION

For this study, we collected the data from the database of a big airline and aircraft aviation. The application is done on number of datasets to compare the results. We are collecting the data between 1970 and 2012. The data is given in a report layout with following sections shown in Table 1. As is apparent from Table 1, component reports have a greater number of attributes. The aim of the analysis is to find the attributes that affect the warning levels and a new formulation about it.

6.1 Application and results

The purpose of these application datasets is to bring out the effective parameters in order to reduce the number of the deaths in the incidents. For the present sample space analysis, WEKA software among other data-mining programs is chosen. In this design we will be using four filters based on feature ranking techniques and one wrapper based feature ranking technique. They are information gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), ReliefF (RFF) and oneR attribute evaluation (OneR). Also, we use the best first methods with Correlation feature subset (CFS) and Consistency Subset (CS). The experiments were performed to evaluate the predictive performance of individual rankers and best first approach. The experiments are also performed on the entire dataset to evaluate the performance between rankers and best first, we have built classification models using k-Nearest Neighbour (KNN), Naive Bayes (NB), Decision Tree (DT), Neural Network (NN) and Support Vector Machines (SVM). The classification models used in our study has ten-fold cross validation as a test method. The proposed classification models are evaluated using the different performance metric like accuracy, specificity and sensitivity.

6.2 Data analysis

Experiments conducted in this study were performed on the datasets collected from large airlines and aviation companies. Aircraft dataset consist of more attributes and number of instances. All the attributes are nominal. The dataset used in the study have one group as training data. With training data the analysis was conducted validity and accuracy of the obtained rules was tested. Briefly, target attribute means whether the result of the incident is safety or not. Table 2 shows the classifiers of accuracy with full dataset before applying feature selection methods. Table 3 shows the description of different feature selection methods from E1 to E8. The accuracy of different classifiers after applying feature selection methods are given in Table 4. The average accuracy of different classifiers is given in Table 5 and the average accuracy of different feature selection methods are given in Table 6.

6.3 Results

It is observed from the comparison of attribute set categorical values, that the results are meaningful. As is apparent from Table 4, classification after applying feature selection methods a result with attributes gives the highest accuracy. It is observed that decision trees are effective in discovering the usable and right information in the data. It is easy for interpreters to understand the results obtained in this work. With the help of the results obtained here, predictions can be made and precautions can be taken against aircraft accidents. As a result, some safety recommendations are addressed to airline aviation administrations. We can say that experienced and certificated pilots should be on duty and commercial pilots should not be assigned to those flights.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

Table 1
Description of sample datasets used in application

Parameter name	Description
Abind	The challenging for the pilot to diagnose in flight
Aflalo	The side of the runway after landing long
Airatt	The aircraft will be put into a round out attitude shortly before it would otherwise contact the ground
Airbrot	The burnt-out aircraft is clustered on the aft section of the flight deck, clear of the fire area
Arspd	The highest airspeed attained by an aircraft of a particular class
Altplm	The standard nominal altitude of an aircraft, in hundreds of feet
Apnopre	The pilot-interpreted make use of ground beacons and aircraft equipment such as VOR, NDB
Arhorfa	The view of heading indicator and artificial horizon after an in flight vacuum failure
ATCerr	The service provided by ground-based controllers who direct aircraft on the ground
Autlad	The designed to make landing possible in visibility too poor to permit any form of visual landing
Autpidiso	The autopilot can control the aircraft while the pilot attends to other duties
Autpieng	The autopilot must be turned on using the autopilot engage switch on the far left
Autthrot	The pilot to control the power setting of an aircraft's engines by specifying a desired flight characteristic
Bdwetr	The atmospheric conditions that comprise the state of the atmosphere in terms of temperature and wind
Belruele	The100ft above the touchdown zone elevation of the runway of intended landing and below that altitude
Brdstr	The bird strikes happen most often during takeoff or landing, or during low altitude flight.
Bomrmnt	The key tactical control and administrative organisation for bombers in all theaters of operation.
Boucd	The pilot lost control of the aircraft landed hard and bounced on the runway during landing.
Brfa	The brakes failed.
Burfla	A light aircraft burst into flames after crashing into a building during an air show.
Cenln	The aircraft exactly on the centerline there will be bumping at each center light.
Cirsf	Inertial reference systems (IRS) use ring laser gyros and accelerometers in order to calculate the aircraft position.
Climb	It is used to climb after takeoff until ready to leave the traffic circuit.
Corctrl	The aircraft structures are made of metal, and the most insidious form of damage to those structures is corrosion.
Crasd	To break in pieces violently; to dash together with noise and violence.
Crswnd	The component of wind that is blowing across the runway making landings and take-offs more difficult
Crstrfa	The structural cruise speed at which speeds must be below to avoid damage in turbulence.
Cvrinfo	It is used to record the audio environment in the flight deck of an aircraft for the purpose of investigation.
Damexpl	Explosions occurred below the aircraft impact levels in both towers prior to the collapses.
Dsnerr	Aircraft design is a compromise between many competing factors and constraints
Dirloctrl	The pilot to control the direction (left or right) of yaw about the aircraft's vertical axis for minor adjustments.
Dsaprd	Many aircraft have gone missing without trace.
Dixmd	The displacement launched
Dme	To determine their distance from a land-based transponder by sending and receiving pulse pairs
Drmalfn	The accident was most likely caused by improper wiring and deficiencies in the door's design
Doendabld	The twin-engine aircraft must be able to fly after losing an engine.
Dwnwnd	The direction of the wind just like blown smoke
Durtkof	For light aircraft, usually full power is used during takeoff
Durtkofr	Usually the engines are run at full power during takeoff
Elfa	The electrical failures often result from interconnection breakdown between aircraft systems.
Emrland	The operation of the aircraft or involves sudden medical emergencies necessitating diversion to the nearest airport
Endruny	That end of the runway nearest to the direction from which the final approach is made.
Enfa	The engine failure is probably your worst fear as a pilot
Enmalfn	The engine failed on the way to town
Eqcolngsyplm	The electrical or electronic equipment on board an aircraft system problem.
Errinstrdng	The instruments in the cockpit of an aircraft that provide the pressure
Foaprd	To act as both an assertive individual and as a subordinate in a team atmosphere.
Fabldmgd	The fan blade damaged
Fagd	The progressive and localised structural damage that occurs when a material is subjected to cyclic loading.
Fctm	The Flight Crew Training Manual provides information and recommendations on manoeuvers and techniques.
Fdrinfo	The performance of various aircraft systems, as well as the aircraft's speed, altitude, heading and flight parameters.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

Fr	The most serious in-flight emergency, and must be brought under control as soon as possible.
Fstmdarcol	A mid-air collision is an aviation accident in which two or more aircraft come into contact during flight
Flplm	When used during takeoff, flaps trade runway distance for climb rate and reduce ground roll and the climb rate.
Fggr	The most frequent cause of surface visibility below three miles.
Fghvy	Heavy fog forced drivers to slow down
Fghl	This type of fog usually forms a good distance from the peak of the hill or mountain and covers a large area.
Fglgt	The fog that obscures less than 60% of the sky and does not extend to the base of any overhead clouds.
Fcdldng	The aircraft is forced to make a landing due to technical problems.
Frmruny	Defined rectangular area on a land aerodrome prepared for the landing and takeoff of aircraft".
Fulex	Exhaustion is when one quits running and falls into a heap, so to is fuel exhaustion with respect to an aircraft
Fulstar	There is a supply problem which either fully or partially prevents the fuel from reaching the engine.
Fultkexpd	Fuel tanks have been implicated in aviation disasters, being the cause of the accident or worsening it.
Fultkigd	Any process or event capable of causing a fire or explosion.
Fultklkd	The easy-to-fix to more difficult and expensive problems.

Table 2
Classifiers accuracy with full dataset before applying feature selection methods

Classifiers	Correctly classified instances	Incorrectly classified instances	Accuracy (%)
DT	422	78	84·4
NB	429	71	85·8
SVM	422	78	84·4
KNN	413	87	82·6
NN	413	87	82·6

Table 3
Description of feature selection methods

Evaluator	Description
E1	Correlation-based feature selection with best first search strategy
E2	Consistency-based feature selection with best first search strategy
E3	Gain ratio feature evaluation with ranker search strategy
E4	Information gain feature evaluation with ranker search strategy
E5	OneR feature evaluation with ranker search strategy
E6	Principal component analysis with ranker search strategy
E7	Relief feature evaluation with ranker search strategy
E8	Symmetric uncertainty feature evaluation with ranker search strategy

7.0 EXPERIMENTAL RESULTS AND DISCUSSION

We have applied the eight feature selection techniques (CFS, CS, GR, IG, OneR, PCA, RFF and SU) to the aircraft dataset. After the feature selection, we used five classifiers DT, NB, SVM, KNN and NN to build classification models on the datasets with various selected subset of features. The classification models are evaluated in terms of the accuracy performance metric.

Table 4 summarise the classification performance in terms of accuracy for different feature selection methods. The table also display model performance on whole dataset. It is observed from the comparison of DT, NB, SVM, KNN and NN in the performance metric of accuracy values, that the results are reliable (see Table 4). Table 5 displays and summarises the comparison of classification

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

Table 4
Performance of different classifers after applying feature selection methods

	E1	E2	E3	E4	E5	E6	E7	E8	AVG
DT	97·08	91·08	99·04	99·04	98·04	99·08	95·04	99·04	97·68%
NB	97·08	90·02	99·04	99·04	98·04	99·08	93·08	99·04	97·28%
SVM	97·08	91·06	99·04	99·04	98·04	99·08	95·04	99·04	97·65%
KNN	97·08	91·02	99·04	99·04	98·00	99·06	94·08	94·04	96·83%
NN	97·08	86·06	99·04	98·04	97·06	99·08	91·04	93·00	95·05%
AVG	97·08%	90·28%	99·04%	99·02%	98·16%	99.76%	94·16%	97·12%	

Table 5
Average accuracy of different classifiers

Classifiers	Accuracy (%)
Decision tree	97·68
Naive Bayes	97·28
Support vector machines	97·65
K-nearest neighbour	96·83
Neural network	95·05

Table 6
Average accuracy of using different feature selection methods

Attribute selection methods	Accuracy (%)
CFS subset evaluation	97·08
Consistency subset evaluation	90·28
Gain ratio attribute evaluation	99·04
Info gain attribute evaluation	99·02
OneR attribute evaluation	98·16
Principal component analysis	99·76
Relief attribute evaluation	94·16
Symmetric uncertainty	97·12

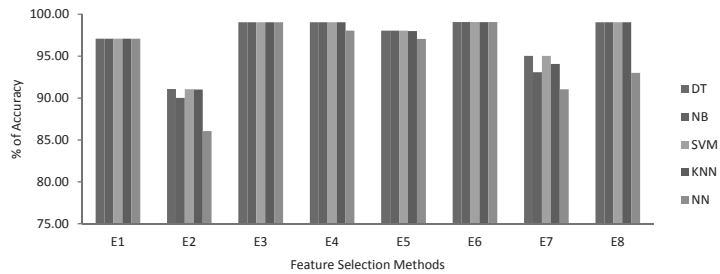


Figure 5. Performance of different classifiers with different feature selectors.

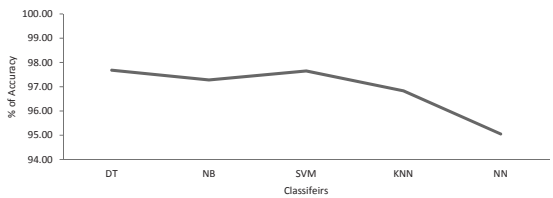


Figure 6. Performance of different classifiers.

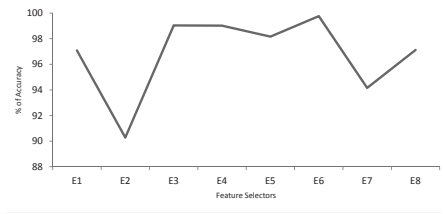


Figure 7. Performance of different feature selection methods.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

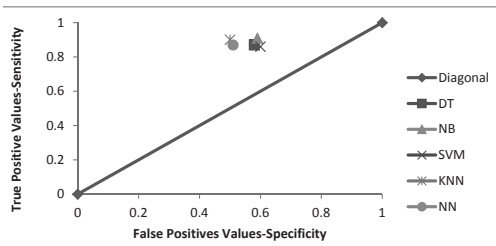


Figure 8. Performance of classification models before applies feature selection with ROC chart.

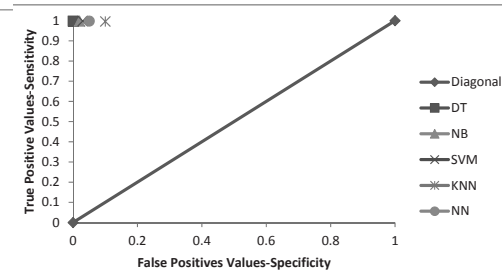


Figure 9. Performance of classification models after applied feature selection (PCA) principal component analysis with ROC chart.

performance in terms of average accuracy of different classifiers. Table 6 shows and summarises the average accuracy of using different feature selection methods.

All these results are mapped into a group of classifiers and features as shown in Figs 5-7. The multiple comparison results are presented in Fig. 5. The figure shows the following facts: DT performed better than other classifiers and PCA methods performed better than other feature selection methods. CFS, CS, GR, OneR, IG, RFF, and SU these seven methods are ordered by their performances from worst to best. The average accuracy of classifiers results presented in Fig. 6. The average accuracy of different feature selection methods results presented in Fig. 7.

The ROC (receiver operating characteristic curve), is a graphical plot which illustrates the performance of a binary classifier as its discrimination threshold is varied. It is created by plotting the fraction of true positives of out of the positives (TPR = True Positive Rate) vs the fraction of false positives out of the negatives (FPR = False Positive Rate), at various threshold settings. TPR is also known as sensitivity (also called recall in some fields), and FPR is one minus the specificity or true negative rate. Figure 8 Performance of classification models before feature selection with ROC chart. Figure 9 Performance of classification models after feature selection (PCA) with ROC chart.

All classifiers produced the analysis of the training dataset and the classification rules. In the experiment, the phase of experiment is the evaluation and interpretation of the classification rules using the unseen data. In the experiment we have used number of instances of the database as a training datasets. The analyses were performed using WEKA environment. This study also proved that the DT classifier would performance better than other classifiers and PCA method would performance better than other feature selection methods on airline data.

8.0 CONCLUSION

In this study, five different classifiers are used for the classification of data. We have explored the use of different classification techniques on aviation accidents data. We have studied the performance of different feature selection methods. We reviewed four filter based feature ranking techniques and one wrapper based feature ranking technique including information gain, gain ratio, symmetrical uncertainty, reliefF and oneR attribute evaluation. The main contribution of this study is to evaluate the performance of different classification techniques are DT, NB, SVM, KNN and NN on aviation components data. In this paper scrutinised the importance of feature selection methods for improving the performance of classification methods. It is found that the different feature selection attribute reduces more number of redundant and irrelevant attributes thereby increases the performance of classifiers. We determine that principal component analysis based with feature selector attributes and decision tree based with classifier is best suitable for aircraft accident data prediction.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press

ACKNOWLEDGEMENT

The authors wish to acknowledge the financial support from the University Grant Commission (UGC), New Delhi, India for the Major Research Project 'Data Tuner for effective Data Pre-processing' vide reference F.No. 39-899/2010 (SR), and also gratefully acknowledge the unanimous reviewers for their kind suggestions and comments for improving this paper.

REFERENCES

1. ARAUZO-AZOFRA, A., BENITEZ, J.M. and CASTRO, J.C. Consistency measures for feature selection, *J Intell Inf Syst*, 2008, **30**, pp 273–292.
2. ASHA, G.K, MANJUNATH, A.S. and JAYARAM, M.A. A comparative study of attribute selection using gain ratio and correlation based feature selection, *Int J of Info Tech and Knowledge Management*, July-December 2012, **2**, pp 271-277.
3. RAMANA, B.V. BABU, M.S.P. and VENKATESWARLU, N.B. A critical study of selected classification algorithms for liver disease diagnosis, *Int J Database Management Systems*, May 2011, **3**, (2).
4. IENCO, D., PENSA, R.G. and MEO, R. *Context-based Distance Learning for Categorical Data Clustering*, 2009, IDA 2009, LNCS 5772, Springer, Berlin, pp 83-94.
5. Recent advances in data mining, *Eng Apps of Artificial Intelligence* (Editorial), **19**, 2006, pp 361-362.
6. GUYON, I. and ELISSEEFF, A. An introduction to variable and feature selection, *J Machine Learning Research*, 2003, **3**, pp 1157-1182.
7. HE, H., GRACO, W. and YAO, X. Application of genetic algorithm and *k*-nearest neighbour method in medical fraud detection, simulated evolution and learning, November 1998, Second Asia-Pacific Conference on Simulated Evolution and Learning, SEAL' 98, Canberra, Australia, Springer, pp 74-81.
8. JIAWEI, H. and KAMBER, M. Data Mining: Concepts and Techniques, 2001, University of Simon Fraser.
9. WANG, H., KHOSHGOFTAAR, T.M. and NAPOLITANO, A. A comparative study of ensemble feature selection techniques for software defect prediction, 2010, Ninth International Conference on Machine Learning and Applications, 12-14 December 2010, Washington, DC, USA, pp 135-140.
10. SHYUR, H.J. A quantitative model for aviation safety risk assessment, *Computers and Ind Eng*, 2007.
11. HAN J. and KAMBER, M. *Data Mining; Concepts and Techniques*, 2000, Morgan Kaufmann Publishers.
12. BINEID, M. and FIELDING, J.P. Development of a civil aircraft dispatch reliability prediction methodology, *Aircraft Eng and Aerospace Tech*, 2003, **75**, (6), 2003, pp 588-594.
13. MURTY, M.N. and DEVI, V.S. *Pattern Recognition: An Algorithmic Approach*, Springer, Ch 4, pp 86-97.
14. GRIMALDI, M., CUNNINGHAM, P. and KOKARAM, A. An Evaluation of Alternative Feature Selection Strategies and Ensemble Techniques for Classifying Music, Trinity College Dublin, Ireland.
15. PIZZI, N.J. and PEDRYCZ, W. *Effective classification using feature selection and fuzzy*, Integration Fuzzy Sets and Systems, 2008.
16. DESSUREAULT, S., SINUHAJI, A. and COLEMAN, P. Data mining mine safety data, *Mining Engineering Littleton*, 2007, **59**, (8), p 64. 7 pgs.
17. PULATOVA, S. *Covering (rule-based) Algorithms Lecture Notes in Data Mining*, 2006, World Scientific Publishing Co, pp 87-97.
18. SOLOMON, S. NGUYEN, H. LIEBOWITZ, J. and AGRESTI, W. Using data mining to improve traffic safety programs, *Ind Management and Data Systems*, 2006, **106**, (5), pp 621–643.
19. SURAJ, Z. and DELIMATA, P. Data mining exploration system for feature selection tasks. In International conference on hybrid information technology, 2006, Ichit'06, *User Manuel of Polyanalyst 5*, April 2005, IEEE, Computer Society.
20. MITCHELL, T. *Machine Learning*, 1997, McGraw Hill.
21. ALTIDOR, W., KHOSHGOFTAAR, T. M., VAN HULSE, J. and NAPOLITANO, A. Ensemble feature ranking methods for data intensive computing applications, *Handbook of Data Intensive Computing*, Springer Science + Business Media, LLC 2011, pp 349-376.
22. CHEN, Y., LI, Y., CHENG, X., GUO, L., LIPMAA, H., YUNG, M. and LIN, D. *Survey and taxonomy of feature selection algorithms in intrusion detection system*, 2006, Inscrypt 2006, LNCS 4318 Springer-Verlag, Berlin, Germany, pp153-167.
23. NAZERI, Z. and JIANPING, Z. Mining aviation data to understand impacts of severe weather on airspace system performance, International Conference on Information Technology, 2002, IEEE.

https://doi.org/10.1017/S0001924000009623 Published online by Cambridge University Press