pp 1849–1866. © Royal Aeronautical Society 2016
doi: 10.1017/aer.2016.107

Large-scale data analysis on aviation accident database using different data mining techniques

A.B. Arockia Christopher
Email: abchristomca@yahoo.com
Department of Information Technology
VSB Engineering College
Karur, Tamil Nadu
India

V. Shunmughavel Vivekanandam
Email: shunsvel@gmail.com
Department of Computer Science and Engineering
VSB Engineering College
Karur, Tamil Nadu
India

A.B. Antony Anderson
Email: tony_br@rediffmail.com
Department of Computer Science and Engineering
PVP College of Technology and Engineering for Women
Dindigul, Tamil Nadu
India

S. Markkandeyan
Email: markkan_s@rediffmail.com
Department of IT, RVS College of Engineering
Dindigul, Tamil Nadu
India

V. Sivakumar
Email:sivamugunthan13@gmail.com
Department of Electrical and Electronics Engineering
VSB Engineering College
Karur, Tamil Nadu
India

Received 22 April 2016; revised 19 July 2016; accepted 11 August 2016.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

ABSTRACT

Data mining is an iterative process in which progress is defined by discovery through either automatic or manual methods. A data cleaning procedure is proposed to improve the quality of classification tasks in the knowledge discovery process by taking into account both redundant and conflicting data. The redundancy check is performed on the original dataset and the resultant dataset is preserved. This resultant dataset is then checked for conflicting data and, if any are found, they are corrected and updated on the original aircraft dataset. This updated dataset is then classified using a variety of classifiers such as Bayes, functions, lazy, MISC, rules and decision trees. The performance of the updated datasets on these classifiers is examine, and the result shows a significant improvement in the classification accuracy after redundancy and conflicts are removed. The conflicts after correction are updated in the original dataset, and when the performance of the classifier is evaluated, great improvement is observed. This paper aims to address how data mining techniques can be used to understand complex system accidents in the aviation domain. Decision trees are considered to be the one of the most powerful and popular approaches in knowledge discovery and data mining. The objective is to develop a classification model for aviation risk investigation and reduction using a decision tree induction method that enhances the ability to form decision trees and thereby proves that the classification accuracy of decision trees is greater. Different feature selectors are used in this study in order to reduce the number of initial attributes.

Keywords: data mining; feature selection; classification techniques; decision tree; aviation

1.0 INTRODUCTION

Data mining is the search for new, valuable and nontrivial information in large volumes of data. It is a cooperative effort by humans and computers. The best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capabilities of computers. In practice, the two primary goals of data mining tend to be prediction and description. Prediction involves using some variables or fields in the dataset to predict unknown or future values of other variables of interest. Description, on the other hand, focuses on finding patterns describing the data that can be interpreted by humans. Therefore, it is possible to put data mining activities into one of two categories: 1) predictive data mining, which produces the model of the system described by the given dataset, or 2) descriptive data mining, which produces new, nontrivial information based on the available dataset. On the predictive end of the spectrum, the goal of data mining is to produce a model, expressed as an executable code, which can be used to perform classification, prediction, estimation or other similar tasks. On the other descriptive end of the spectrum, the goal is to gain an understanding of the analysed system by uncovering pattern and relationships in large scale aircraft dataset.

Feature selection is a term commonly used in data mining to describe the tools and techniques available for reducing inputs to a manageable size for processing and analysis. Feature selection implies not only cardinality reduction, which means imposing an arbitrary or predefined cut-off on the number of attributes that can be considered when building a model, but also the choice of attributes, meaning that either the analyst or the modelling tool actively selects or discards attributes based on their usefulness for analysis. The ability to apply feature selection is critical for effective analysis, because datasets frequently contain far more information than is needed to build the model. Even if resources are not an issue,

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

one typically wants to remove unwanted features because they might degrade the quality of discovered patterns because some features are noisy or redundant. This noise makes it more difficult to discover meaningful patterns in the data. To discover quality patterns, the data mining algorithms require much larger training aircraft accident dataset on large scale dataset.

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 symmetrical uncertainty attribute.

In this paper, we applied different data mining methods on the aircraft accident reports. We found that feature selection between attributes and classification algorithms to find the relations and rules about the accidents resulted in fatality. We also used Weka as a pre-processing tool to select the attributes to be used in the large-scale data 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 are used to predict the warning level of the component as the class attribute.

We have different classifier algorithms such as Bayes classifiers, functions classifiers, lazy classifiers, MISC classifiers, rules classifiers, and decision tree classifiers to be considered for comparing their performance based on the large-scale aircraft dataset. In this experimentation, tenfold cross-validation methods for comparison of selected classification algorithms have been used. This paper concentrates on performance of classifier algorithms with different features combinations such as CFS, consistency, GR, information gain, OneR, PCA, ReliefF and symmetric with the selected datasets and on the improvement of the classification accuracy.

The paper is organized as follows. In Sections 2 and 3, we discuss the related work and proposed work. In Section 4, we produce the experimental results and discussion about classification with feature selection rule discovery for the large-scale aviation dataset.

2.0 RELATED WORK

Data mining is a multidisciplinary effort to extract nuggets of knowledge from data. The proliferation of large datasets within many domains poses unprecedented challenges to data mining as defined by Han and Kamber[14]. Not only are datasets getting larger, but new types of data are also becoming prevalent, such as data streams on the web, microarrays in genomics and proteomics, and networks in social computing and systems biology. Researchers are realizing that in order to achieve successful data mining, feature selection is an indispensable component, as proposed by Guyon and Elisseeff[13] and Liu and Motoda[17].

Feature selection is normally done by searching the space of attribute subsets and evaluating each one. This is achieved by combining an attribute subset evaluator with a search method. In the present investigation, an evaluation of six filter feature subset methods with a rank search or greedy search method was performed to find out the best feature sets, and they are listed below for reference: (1) correlation-based attribute evaluation (CB); (2) chi-square attribute evaluation (CH); (3) Gain-Ratio attribute evaluation (GR); (4) Information Gain attribute evaluation (IG); (5) ReliefF attribute evaluation (RF); and (6) Symmetrical Uncertainty attribute evaluation (SU). That these entire filter techniques mentioned above could assess the relevance of features has been studied by Saeys et al[31] on the basis of the intrinsic properties

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

of the data. Feature selection often increases classifier efficiency by reducing the size of the effective features.

Correlation feature subset selection 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 a decrease with growing inter-correlation as has been studied by Asha et al[5]. The study by Antonio et al[3] established that, in order to predict the concept or class value of its instances, a dataset with the selected features alone must be consistent. 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 latter is more informative, as defined by Asha et al.

Information Gain (IG) 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, according to the method proposed by Altidor et al[2]. Rule-based algorithms provide ways to generate compact, easy-to-interpret and accurate rules by concentrating on a specific class. One way of generating classification rules is to use decision trees. The disadvantage of using a decision tree is its complex and incomprehensible nature, which has been noted in the literature[25]. 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 and has been studied by Grimaldi et al[18].

ReliefF is an easy-to-use, fast and accurate algorithm even with dependent features and noisy data as defined by Wang et al[29]. Symmetric uncertainty is used to measure the degree of association between discrete features. It is derived from entropy that has been proposed by Chen et al[7]. 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, indicating that one variable (either X or Y) completely predicts the other variable, has been studied by Ienco et al[15].

Feature selection is a process of selecting a subset of original features according to certain criteria and is an important and frequently used technique in data mining for dimension reduction. It reduces the number of features; removes irrelevant, redundant or noisy features; and brings about palpable effects for applications, thus speeding up a data mining algorithm, improving learning accuracy and leading to better model comprehensibility. Various studies show that some features can be removed without performance deterioration as shown by Ng[23] and Donoho[9].

Feature selection can be found in many areas of data mining such as classification, clustering, association rules and regression. For example, feature selection is called subset or variable selection in statistics[19]. Guyon and Elisseeff[13] have proposed the reducing of the number of dimensions by selecting variables or features which is effective in dealing with high-dimensional data. Variable and feature selection have become the focus of much research in areas of application for which datasets with many number 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. Feature selection has been an active research area in the pattern recognition, statistics and data mining communities.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

A decision based on selected features about whether a dangerous surgery is necessary has been studied by Kim et al[30]. Delimata and Suraj[28] have dealt with feature selection, which is usually meant as a process of finding a subset of features from the original set of features forming patterns in a given dataset, optimised according to the defined goal and criterion of feature selection. Many classification problems involve features whose specificity demands some form of feature space transformation (pre-processing) coupled with post-processing consensus analysis in order to accomplish a successful discrimination between different classes. Pizzi and Pedrycz[24] have presented a new methodology that systematically addresses these design classification issues in their study. At the pre-processing phase, they offer a new approach of stochastic feature selection. This type of feature selection collates quadratically transformed feature subsets for presentation to a collection of respective classifiers. Feature selection has been an active and fruitful field of research area in the pattern recognition, machine learning, statistics and data mining communities and has been studied by Mitra et al[20] and Miller[19].

Subset generation is a search process that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation. If the new subset turns to be better, it replaces the best one. This process is repeated until a given stopping condition is satisfied. Ranking of features determines the importance of any individual feature, neglecting their possible interactions. Ranking methods are based on statistics, information theory or on some functions of the classifier's outputs, as studied by Duch et al[10].

Decision tree induction as defined by Han and Kamber[14] is a very popular and practical approach for pattern classification. Narasimha and Susheela[21] have dealt with the Naïve Bayes (NB) classifier, a simple probabilistic classifier based on the Bayes theorem where every feature is assumed to be class-conditionally independent. 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. A Support Vector Machine (SVM) is a hyperplane that separates two different sets of samples with maximum distance of hyperplane to the nearest samples from both sets as by Han and Kamber[14]. 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 neighbours as described by Han and Kamber.

The automatic and smart analysis of the complex structure of high-volume data in the aviation industry has been studied by Gürbüz et al[11]. A predictive model to explain the development of a dispatch reliability prediction method for passenger aircraft was proposed by Bineid and Fielding[6]. Nazeri and Zhang[22] have described the application of data mining to analysing severe weather impacts on National Airspace System (NAS). A model that allows investigation of non-linear effects of aviation safety factors and flexible assessment of aviation risk has been described in the literature[26]. The main goal of this work 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 data mining analysis method that can be applied to evaluate that monitor red-light-signal controlled intersections to improve traffic safety by reducing fatalities has been demonstrated by Solomon et al[27].

Considering the high level of interaction between a service provider and its customers, the significance of a marketing approach cannot be overstated[4]. Over the past few decades, considerable attention has been focused on cultivating a market-driven approach, which fosters the effective implementation of the marketing concept as an important source of competitive advantage in the tourism industry. Nazeri and Zhang[22] have defined data

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

mining methods that have been successfully introduced in many fields. It is still a research topic, but of tremendous interest to the industry in order to solve its real-world problems. Statistics, artificial intelligence techniques, decision tree approaches, genetic algorithms and visualization 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 or accidents have also dramatically increased in recent years.

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. Nowadays, 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 aviation industry tools with capable instruments are needed.

Data mining is a potentially useful instruments that was not known before, and using this instrument to reveal information hidden in the data was proposed by Jiawei and Kamber[16]. As a result, there is a large amount of knowledge and data accumulation in the aviation industry. These 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 have been used.

Bineid and Fielding[6] have used data mining techniques to explain the development of a dispatch reliability prediction method for passenger aircraft. The U.S. 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 and help guide research, and best practices have been studied by Dessureault et al[8].

Dessureault et al have explored the background of the Part 50 database, given a general background of data warehousing and data mining, and presented some of the interesting analyses that resulted from a modernized Part 50 data warehouse using data mining. 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 to an aviation accident. The potential human-error-related risk could be identified and monitored in a timely fashion. The results can provide better information to the civil aviation communities to manage aviation safety risk so corrective action can be taken to reduce the occurrence of aviation accidents. It also allows identification of the fundamental causes 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 as described by Shyur[26].

This work represented several data modelling techniques such as 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 relationships 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 U.S. Department of Transportation from fatal accidents at red-light-signal

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

controlled intersections in Maryland and Washington, DC, from 2000–2003 by Solomon et al[27].

3.0 PROPOSED METHOD

Data mining is a process of discovering various models, summaries and derived values from a given collection of large-scale aircraft accident data. It is important to realize that the problem of discovering or estimating dependencies from data or discovering totally new data is only one part of the general experimental procedure used by scientists, engineers and others who apply standard steps to draw conclusions from large-scale aircraft data. The general experimental procedure adapted to data mining problems involves the following steps.

1. Statement of the problem
2. Collection of aircraft accident data
3. Pre-processing of aircraft accident data
4. Estimate of the model (tenfold cross-validation model)
5. Interpretation of the model and formulation of conclusions

Steps 3 and 4 in the data mining process have to be understood just as two steps in a more complex process. All phases separately and the entire data mining process as a whole are highly iterative, as is shown in Fig. 1. A good understanding of the whole process is important for any successful application. No matter how powerful the data mining method used in Step 4 is, the resulting model will not be valid if the data are not collected and pre-processed correctly, or if the problem formulation is not meaningful. In general, feature selection works by calculating a score for each attribute and then selecting only the attributes that have the best scores. We can also adjust the threshold for the top scores. Analysis services provide multiple methods for calculating these scores, and the exact method that is applied in any model depends on these factors: (1) the algorithm used in one's model, (2) the data type of the attribute, and (3) any parameters one has set on the model.

Feature selection is applied to inputs, predictable attributes or states in a column. When scoring for feature selection is complete, only the attributes and states that the algorithm selects are included in the model-building process and thus can be used for prediction. If one chooses a predictable attribute that does not meet the threshold for feature selection, the attribute can still be used for prediction, but the predictions will be based solely on the global statistics that exist in the model.

A feature selection algorithm can be seen as the combination of a search technique for proposing new feature subsets along with an evaluation measure which scores the different feature subsets. The simplest algorithm is to test each possible subset of features and finding the one which minimizes the error rate. A typical feature selection process consists of four basic steps shown in Fig. 2: subset generation, subset evaluation, stopping criterion and result validation.

Decision tree induction is a very popular and practical approach for pattern classification. A 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. A 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

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

1. Statement of the problem
2. Collection of the aircraft accident data
3. Pre-processing of the aircraft accident data
4. Estimate of the model (10 fold cross validation model)
5. Interpretation of the model and drawing conclusions

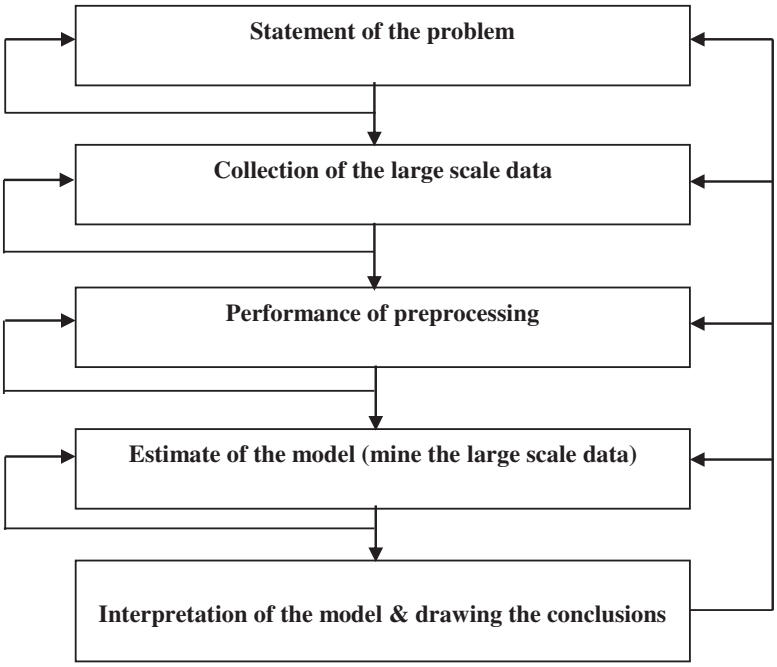


Figure 1. The data mining process.

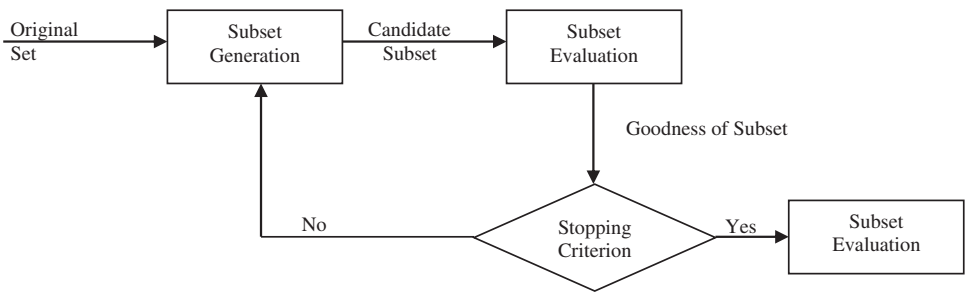


Figure 2. Four key steps of feature selection.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

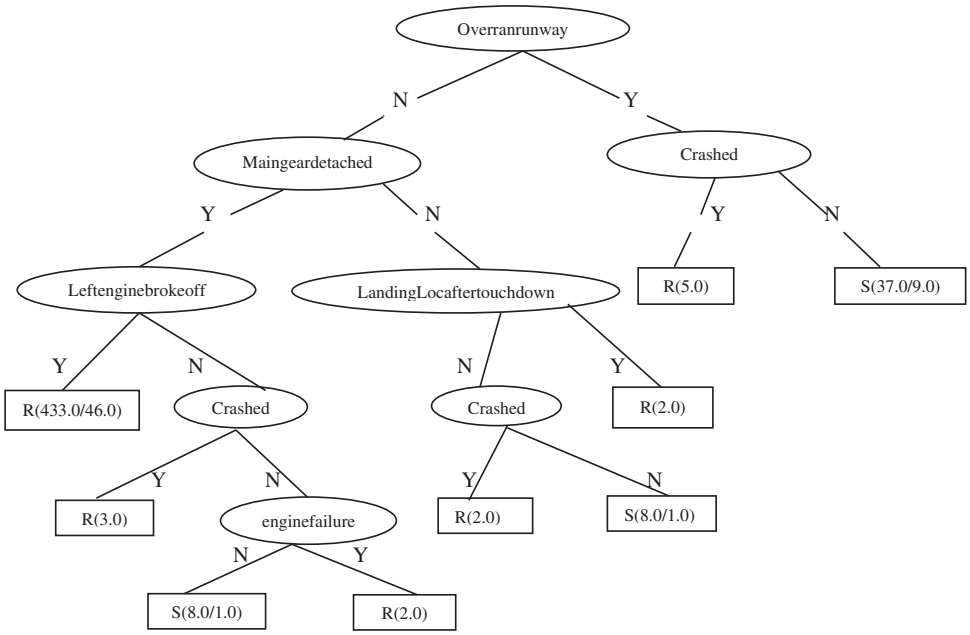


Figure 3. The decision tree model for an aircraft accident dataset.

unknown value for class label. the decision tree is constructed in two phases: the building phase and the 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, the tree is constructed with that as the root node and the distinct values of the attribute are denoted as branches. The process of selecting the best attribute and representing the distinct values as branches is 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 aircraft accident dataset. 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 and are easy to understand. Construction of decision trees is fast and yields better accuracy. There are several algorithms to classify data using decision trees.

J48 s are the improved versions of C4.5 algorithms. or they can be called an optimized implementation of the C4.5 s. The output of J48 is the decision tree. A decision tree is similar to a tree structure with a root node, intermediate nodes and leaf nodes. Each node in the tree consists of a decision that leads to a result. Decision trees divide the input space of a dataset into mutually exclusive areas, with each area having a label, value or action to describe its data points. A splitting criterion is used to calculate which attribute is the best for splitting that portion tree of the training data that reaches a particular node. Figure 3 shows the decision tree model using J48 for an aircraft dataset, whether there was incident or not. In this model, R means Risk Factor and S means Safety for decision tree formed by using the accident attribute of the aircraft dataset.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

Input: The training samples, attribute-list, Data in .arff format.

Output: Prediction of training data.

Method:

1) Input is read in the .arff format and stored into the database.
2) Feature Selection Methods using attribute is done.
3) Eliminating irrelevant attributes (A_1, A_2, A_3…, A_n) are identified.
4) Reduced attributes set (A_1, A_2, A_3…, A_n) are after eliminate.
5) Evaluating the classification algorithms.
6) The class value found, CL_j is updated in the original training data.
7) The Conflicting (C_1, C_2, C_3…, C_n) and redundant data (R_1, R_2, R_3…, R_n) found are also removed.
8) Performance analysis for the predicted data is done for various classifiers using various evaluation techniques (10-fold validation and using training set).

Figure 4. Basic algorithm for prediction (FSC) from training samples.

We applied different classification algorithms for aircraft dataset such as DT, NB, SVM, KNN and NN. First we applied these algorithms for entire dataset. The aircraft Attribute Relation File Format (ARFF) will contain a large quantity of data, and applying classification algorithms to this dataset is time-consuming and also gives less accurate results. Hence we have to reduce the number of initial attributes by using different feature attribute selection methods. After the eliminating irrelevant attributes (A1, A2, A3, …, An) and reduced attributes set (A1, A2, A3, …, An) 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 applied for the aircraft dataset. From that, we concluded that classification accuracy after applying feature selection attribute methods is significantly improved in this model. The Feature Selection Classification (FSC) algorithm is summarized in Fig. 4. The main steps of the proposed FSC algorithm are outlined as follows.

Step 1: Give a new aircraft dataset as input.

Step 2: Apply feature selection methods to initial data.

Step 3: Eliminate irrelevant feature attributes.

Step 4: Reduce attributes of the set after elimination.

Step 5: Evaluate the classifier's performance.

Step 6: Update the class value found as "risk" or "safety" in the original training data.

Step 7: Remove conflicting and redundant data.

Step 8: Analyse the data using various classification algorithms.

This model is proposed for improving classification accuracy by combining the prediction of multiple classifiers. The different data mining classifiers such as Bayes, functions, lazy, MISC, rules and decision trees are used for classification. This model uses tenfold cross-validation as a test method. The classification models are evaluated using different

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

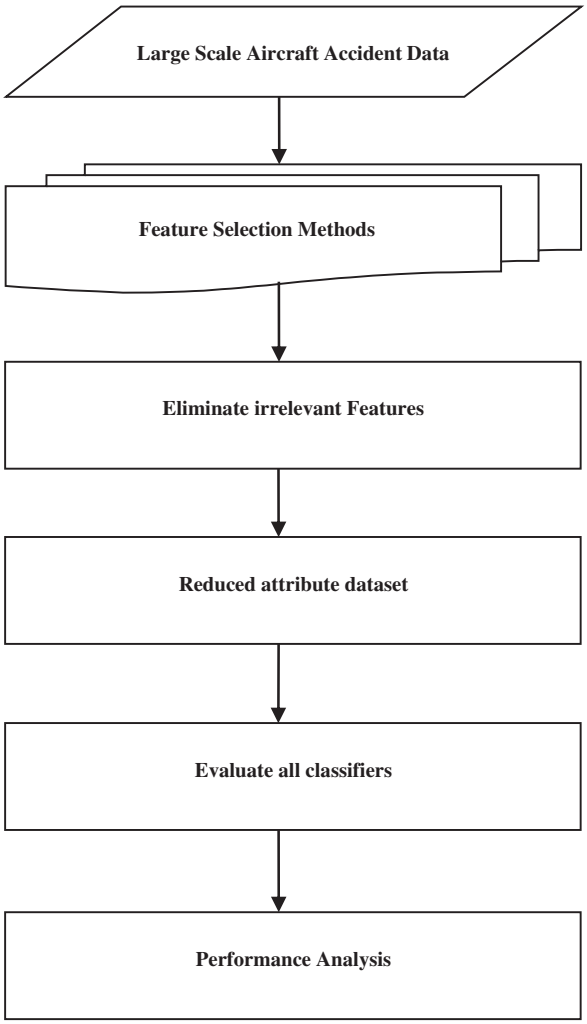


Figure 5. Block diagram for prediction (FSC) model.

performance metrics. A comparative study is carried out 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. We proposed a different feature selection method in this model. The number of features obtained in the dataset is very large. We have 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.

This model found that the feature selection method attribute reduces a greater number of redundant and irrelevant attributes, thereby increasing the performance of classification methods. Figure 5 displays the block diagram for prediction (FSC) model of training samples.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

Table 1
Performance of different Bayes classifiers on reduced features by different feature selectors

Classifiers	CFS	CSY	GR	IG	OneR	PCA	RF	SU
Bayes								
AODE	93.251	93.603	93.642	93.740	93.270	93.349	93.172	93.818
AODEsr	93.251	93.740	93.681	93.798	93.270	93.212	93.172	93.877
BayesNet	93.270	93.446	93.662	93.681	93.212	93.290	93.094	93.759
HNB	93.251	94.229	93.740	93.622	93.231	93.642	93.231	93.838
NaiveBayes	93.270	93.466	93.662	93.681	93.212	93.388	93.094	93.759
NBSimple	93.270	93.466	93.662	93.681	93.212	93.388	93.094	93.759
NBUpdateable	93.270	93.466	93.662	93.681	93.212	93.388	93.094	93.759
WAODE	93.231	94.229	93.759	93.838	93.172	93.857	93.251	93.759

Table 2
Performance of different function classifiers on reduced features by different feature selectors

Classifiers Functions	CFS	CSY	GR	IG	OneR	PCA	RF	SU
Logistic	93.153	95.852	93.720	93.740	92.996	93.309	95.070	93.877
Multilayer	93.192	94.131	93.701	93.427	93.309	92.957	94.972	93.427

In this diagram, we can eliminate irrelevant features and evaluate all classifiers, thereby increasing the performance of classification accuracy and the efficiency of this model. All classifier techniques and different feature selection methods are applied in this model for large-scale dataset analysis.

4.0 EXPERIMENTAL RESULTS AND DISCUSSION

The main contribution here is the study of the classification performance in terms of accuracy with different feature selection methods. A comparison of different classifiers based on the performance metrics (TP rate, FP rate, precision, recall, F-measure, kappa statistic, means absolute error and root mean squared error) on reduced features by different feature selection methods also found that the results are reliable.

Table 1 shows the performance of different Bayes classifiers such as Aode, Aodesr, BayesNet, HNB, naïve Bayes, naïve Bayes simple and naïve Bayes updatable on an aviation large-scale dataset. Table 2 shows the performance of different function classifiers with different feature selection methods. Table 3 shows the performance different lazy classifiers with different feature selection methods. Table 4 shows the performance of different MISC classifiers with different feature selection methods. Table 5 shows the performance different rules classifiers with different feature selection methods. Table 6 shows the performance different decision tree classifiers with different feature selection methods.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

Table 3
Performance of different lazy classifiers on reduced features by different feature selectors

Classifiers Lazy	CFS	CSY	GR	IG	OneR	PCA	RF	SU
IB1	88.693	91.881	87.147	90.395	92.859	90.414	92.351	89.886
IBk	93.133	93.720	93.720	93.388	93.446	92.996	94.718	93.720
KStar	93.016	94.268	92.742	93.251	92.938	93.368	94.835	93.329

Table 4
Performance of different MISC classifiers on reduced features by different feature selectors

Classifiers MISC	CFS	CSY	GR	IG	OneR	PCA	RF	SU
HyperPipe	76.995	76.936	77.034	76.838	77.073	76.643	76.486	76.838
VFI	89.064	93.798	88.497	92.331	92.546	92.214	94.287	91.979

Table 5
Performance of different rules classifiers on reduced features by different feature selectors

Classifiers Rules	CFS	CSY	GR	IG	OneR	PCA	RF	SU
Conjunctive Rule	92.859	92.859	92.859	92.859	92.859	92.859	92.859	92.859
DecisionTable	93.016	93.114	93.270	93.407	93.270	93.036	93.251	93.485

The performance comparison of different Bayes classifiers, functions classifiers, lazy classifiers, MISC classifiers, rules classifiers and decision tree classifiers are shown in Tables 7–12, respectively.

Table 13 shows the performance of accuracy with minimum time (in seconds) taken for different classifiers on an aviation accident large-scale dataset. Decision Tree (J48) will be considered in this large-scale aviation dataset with high accuracy and minimum time (seconds) taken. In this table, both function (logistics) and decision tree (J48) classifiers have accuracy greater than 95%, but function classifiers will take a longer time to execute. Thus, we find that the decision tree classifier will perform better than other classifiers on an aircraft accident dataset with an accuracy of 95.266% with a minimum of time taken. The decision tree classifier accuracy and efficiency of the performance will therefore be increased. Figure 6 displays the accuracy of different classifiers on an aviation large-scale dataset. The decision tree classifier is better than other classifiers because of its high accuracy and minimum time required for execution.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

Table 6
Performance of different decision tree classifiers on reduced features by different feature selectors

Classifiers Decision Trees	CFS	CSY	GR	IG	OneR	PCA	RF	SU
ADTree	92.879	92.899	93.114	92.899	92.996	92.899	92.899	92.899
DecisionStump	92.859	92.859	92.859	92.859	92.859	92.859	92.859	92.859
FT	93.075	95.422	93.720	93.701	93.309	94.307	94.972	93.798
Id3	93.153	93.916	93.642	93.427	93.446	92.899	94.874	93.740
J48	93.075	95.324	93.427	93.564	93.114	93.075	94.933	93.662
J48graft	93.094	95.403	93.544	93.642	93.114	93.603	94.953	93.759
LADTree	93.055	92.820	93.251	92.977	93.153	93.094	92.820	92.977
RandomForest	93.192	94.698	93.701	93.603	93.446	93.642	94.737	93.838
RandomTree	93.153	94.698	93.622	93.505	93.388	92.507	94.777	93.740
REPTree	92.996	95.148	93.427	93.427	93.075	93.290	94.737	93.662
UserClassifier	76.486	76.486	76.486	76.486	76.486	76.486	76.486	76.486

Table 7
Performance of different Bayes classifiers on an aviation accident large-scale dataset

Classification Bayes	Average Accuracy (%)	Time (Seconds) Taken
AODE	93.8381	0.97
AODEsr	93.7989	2.08
BayesNet	93.7793	0.23
HNB	94.3271	1.68
NaiveBayes	93.7402	0.11
NaiveBayesSimple	93.7402	0.08
NaiveBayesUpdateable	93.7402	0.03
WAODE	94.4249	2.47

Table 8
Performance of different function classifiers on aviation accident large-scale dataset

Classification Functions	Average Accuracy (%)	Time (Seconds) Taken
Logistic	95.6768	8.52
MultilayerPerceptron	89.7646	2446.42

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

Table 9
Performance of different lazy classifiers on an aviation accident large-scale dataset

Classification Lazy	Average Accuracy (%)	Time (Seconds) Taken
IB1	91.2167	0.02
IBk	93.3685	0.00
KStar	93.6228	0.00

Table 10
Performance of different MISC classifiers on an aviation accident large-scale dataset

Classification MISC	Average Accuracy (%)	Time (Seconds) Taken
HyperPipes	76.0644	0.03
VFI	94.5915	0.05

Table 11
Performance of different rules classifiers on an aviation accident large-scale dataset

Classification Rules	Average Accuracy (%)	Time (Seconds) Taken
ConjunctiveRule	92.8599	0.84
DecisionTable	93.1142	219.95

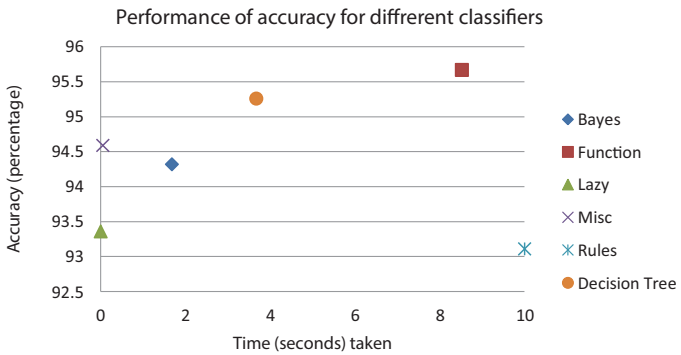


Figure 6. (Colour online) Performance of accuracy with minimum time execution for different classifiers on an aviation large-scale dataset.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

Table 12
Performance of different decision tree classifiers on an aviation accident large-scale dataset

Classification Decision Trees	Average Accuracy (%)	Time (Seconds) Taken
ADTree	92.8991	3.11
DecisionStump	92.8599	0.3
FT	95.0704	38.97
Id3	93.4859	2.54
J48	**95.266**	**3.67**
J48graft	95.3638	7.24
LADTree	92.8208	14.84
RandomForest	94.8748	3.3
RandomTree	92.9773	0.27
REPTree	95.1095	12.06
UserClassifier	76.4867	30.5

Table 13
Performance of accuracy for different classifiers on an aviation accident large-scale dataset

Classifiers	Accuracy (%)	Time (Seconds) taken
Bayes (WAODE)	94.32	1.68
Function (Logistics)	95.67	8.52
Lazy (KStar)	93.36	0.00
MISC (VF1)	94.59	0.05
Rules (Decision Table)	93.11	219.95
Decision Tree (J48)	**95.26**	**3.67**

5.0 CONCLUSIONS

In this study, we have explored the use of different feature selection methods and classification techniques on aviation accident large-scale datasets. We have studied the performance of different feature selection methods and classification algorithms. The main contribution of this study is to evaluate the performance of different feature selection methods such as information gain, gain ratio, symmetrical uncertainty, ReliefF and OneR attribute evaluation on aviation component data. We found that the different feature selection methods reduce more redundant and irrelevant attributes and thereby improve the performance of classifiers. We found that the decision tree classifier proved to be better than other classification techniques by having better classification accuracy and lower misclassification rates. The proposed work is used to predict a component's warning level in aircraft accidents. The rules that were found by this analysis will be helpful for experts in the aviation industry to provide safety recommendations to airline companies.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

ACKNOWLEDGEMENT

The authors wish to acknowledge the anonymous reviewers for their kind suggestions and comments for improving this paper.

REFERENCES

1. Aircraft accident database dataset retrieved from the website: http//www.planecrashinfo.com/database.htm.
2. 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*, 2011, Springer Science Business Media LLC, New York, New York, US, pp 349-376.
3. ANTONIO, A.-A., BENITEZ, J.M. and CASTRO, J.L. Consistency measures for feature selection, *J. Intelligent Information Systems*, 2008, **30**, (3), pp 273-292.
4. APPIAH-ADU, K., FYALL, A. and SINGH, S. Marketing culture and customer retention in the tourism industry, *Services Industries J*, 2000, **20**, (2), pp 95-114.
5. 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. Information Technology and Knowledge Management*, 2012, **2**, pp 271-277.
6. BINEID, M. and FIELDING, J.P. Development of a civil aircraft dispatch reliability prediction methodology, *Aircr Engineering and Aerospace Technology*, 2003, **75**, (6), pp 588-594.
7. 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, *Inscrypt*, 2006, LNCS 4318, Springer-Verlag, Berlin, Germany, pp 153-167.
8. DESSUREAULT, S, SINUHAJI, A. and COLEMAN, P. Data mining mine safety data, *Mining Engineering*, 2007, **59**, (8), pp 64.
9. DONOHO, D. For most large underdetermined systems of linear equations the minimal l^1-norm near-solution is also the sparsest solution, *Communications on Pure and Applied Mathematics*, 2006, **59**, pp 907-934.
10. DUCH, W., WINIARSKI, T., BIESIADA, J. and KACHEL, A. Feature ranking, selection and discretization, International Conference on Artificial Neural Networks (ICANN) and Int. Conf. on Neural Information Processing (ICONIP), 2003, 251-254.
11. GÜRBÜZ, F., ÖZBAKIR, L. and YAPICI, H. Classification rule discovery for the aviation incidents resulted in fatality, *Knowledge-Based Systems*, 2009, **22**, pp 622-632.
12. GÜRBÜZ, F., ÖZBAKIR, L. and YAPICI, H. Data mining and preprocessing application on component reports of an airline company in Turkey, *Expert Systems with Applications*, 2011, **38**, pp 6618-6626.
13. GUYON, I. and ELISSEEFF, A. An introduction to variable and feature selection, *J Machine Learning Research*, 2003, pp 1157-1182.
14. HAN, J. and KAMBER, M. *Data Mining: Concepts and Techniques*, 2001, Morgan Kaufmann Publishers Inc., San Francisco, California, US.
15. IENCO, D., PENSA, R.G. and MEO, R. *Context-Based Distance Learning for Categorical Data Clustering*, IDA 2009, LNCS 5772, Springer, Berlin, Germany, pp 83-94.
16. JIAWEI, H. and KAMBER, M. *Data Mining: Concepts and Techniques,* 2011, Morgan Kaufmann Publishers Inc., San Francisco, California, US.
17. LIU, H. and MOTODA, H. *Computational Methods of Feature Selection*, 2007, Chapman and Hall/CRC Press, Boca Raton, Florida, US.
18. GRIMALDI, G., CUNNINGHAM, P. and KOKARAM, A. An evaluation of alternative feature selection strategies and ensemble techniques for classifying music, Workshop on Multimedia Discovery and Mining, Ireland, 2003.
19. MILLER, A. *Subset Selection in Regression*, 2nd ed, 2002, Chapman and Hall/CRC Press, Boca Raton, Florida, US.
20. MITRA, P., MURTHY, C.A. and PAL, S.K. Unsupervised feature selection using feature similarity, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, **24**, (3), pp 301-312.
21. NARASIMHA, M. and DEVI, V.S. *Pattern Recognition: An Algorithmic Approach*, 2011, Springer, pp 86-97.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press

22. NAZERI, Z. and ZHANG, J. Mining aviation data to understand impacts of severe weather on airspace system performance, Procedings of the International Conference on Information Technology, Las Vegas, Nevada, US, 2002.
23. NG, A.Y. Feature selection, l1 vs. l2 regularization, and rotational invariance, 21st International Conference on Machine Learning, 2004, ACM, New York, New York, US.
24. PIZZI, N.J. and PEDRYCZ, W. Effective classification using feature selection and fuzzy integration, *Fuzzy Sets and Systems*, 2008.
25. BERRY, M. AND BROWNE, M. *Lecture Notes in Data Mining*, 2006, World Scientific Publishing Co., Singapore, pp 87-97.
26. SHYUR, H.J. A quantitative model for aviation safety risk assessment, *Computers and Industrial Engineering*, 2008, **54**, (1), pp 34-44.
27. SOLOMON, S., NGUYEN, H., LIEBOWITZ, J. and AGRESTI, W. Using data mining to improve traffic safety programs, *Industrial Management and Data Systems*, 2006, **106**, (5), pp 621-643.
28. DELIMATA, P. and SURAJ, Z. Data mining exploration system for feature selection tasks, International Conference on Hybrid Information Technology, November 2006, Cheju Island, Korea.
29. WANG, H., KHOSHGOFTAAR, T.M. and NAPOLITANO, A. A comparative study of ensemble feature selection techniques for software defect prediction, *Proc of the Ninth International Conference on Machine Learning and Applications*, 2010, Washington, DC, US, pp 135-140.
30. KIM, Y., STREET, W.N. and MENCZER, F. *Feature Selection in Data Mining*, 2003, IGI Publishing, Hershey, Pennsylvania, US.
31. SAEYS, Y., INZA, I. and LARRAÑAGA, P. A review of feature selection techniques in bioinformatics, *Bioinformatics*, 2007, **23**, (19), pp 2507-2517.

https://doi.org/10.1017/aer.2016.107 Published online by Cambridge University Press