Research on the Data Exploration and Classification Algorithms for Aircraft Safety Accidents

|  |  |
| --- | --- |
| Kai-Jung Huang | Mu-Yen Chen |
| *Department of Information Management* | *Department of Information Management* |
| *National Taichung University of Science and Technology* | *National Taichung University of Science and Technology* |
| *129, Sec 3, Sanmin Rd, Taichung City 404, Taiwan ROC* | *129, Sec 3, Sanmin Rd, Taichung City 404, Taiwan ROC* |
| s1810831013@nutc.edu.tw | mychen@nutc.edu.tw |
| Min-Hsuan Fan |
| Department of Information Management |
| National Taichung University of Science and Technology |
| 129, Sec 3, Sanmin Rd, Taichung City 404, Taiwan ROC |
| mfan@nutc.edu.tw |

***Abstract*** - With the development of air transportation, improving the human ability to explore the surroundings and the world, it has also expanded economic development and various commercial activity. When a serious accident occurs, the first impact is the public’s willingness to take the plane and reduce business activities, may even cause economic recession. In the past, many scholars have also devoted themselves to the research of aviation safety, most of them focus on feature extraction and selection, the differences in accuracy between various classification algorithms are rarely discussed, so this article will compare the results of various classification algorithms, trying to find the most suitable classification algorithm for aviation safety. In the stage of data pre-processing in this experiment , in order to deal with the problems of data imbalance (the difference between the number of casualties and non-fatal accidents), the Synthetic Minority Oversampling Technique (SMOTE) technology will be used to balance the amount of data, then use eight classification algorithms to classify each (using 10-fold cross-validation) for training and testing, finally, this paper take their own evaluation indicators (precision, F-measure, AUC, Recall) to compare, hope after comparing with each other can discover the different characteristics of each method, and further explore the selected attributes.

***Keywords*** - Flight safety, classification algorithms, data preprocessing

**I. Introduction**

According to International Air Transport Association (IATA) 2018’s Flight Safety Report, the average air passenger loss rate per million flights is 0.19, that is, one serious accident will occur in 5.4 million. If you can use the historical data of the Flight Safety accident to carry out exploration, maybe you can find some key causes of casualties and prevent them to reduce the risk of accidents. Because flight safety is very important, how to reduce flight risks has been widely studied.

The data source used in this research is the Accident/Incident Data System (AIDS) provided by the Federal Aviation Administration (FAA). In this data set, all accidents occurred due to flight accidents, but most accidents can be resolved, for example, a slight damage to the body will not cause a sharp drop. However, few of accidents will still cause serious casualties, therefore, the goal of this study is to find out the key factors causing casualties after the flight accident.

The data set used in this research is an example of choosing the more common Boeing model. This data set has a problem that it is imbalanced, that is, the number of casualty and non- casualty data is very different. To balance the effect of data imbalance on the results of classification in fewer categories, this study uses a method called Synthetic Minority Oversampling Technique (SMOTE) to increase the sample of less categories and use Information Gain (IG) to filter characteristic attribute. In the part of classification modeling, this study uses nine classification methods for comparison such as Naive Bayes, Bayes Net, Support Vector Machine (SVM), C4.5, Radial Basis Function Network (RBFN), etc. At last, this study uses the Confusion Matrix as the main performance indicator.

1. Find out which characteristics are likely to cause casualties

2. Apply Synthetic Minority Oversampling Technique sample-increasing technology to solve the problem of

unbalanced data category

3. Apply information gain to sort attributes and use threshold to filter feature attributes。

**II. Literature Review**

2.1 Synthetic Minority Oversampling Technique

Synthetic Minority Oversampling Technique (SMOTE) is a synthetic minority oversampling technique (Chawla et al., 2002). This algorithm is based on improve Random Oversampling algorithm, because Random Oversampling only simply copy to increase a few category samples will cause Over Fitting. The basic method of ​​the SMOTE algorithm is to use the characteristics of all the minority class samples to rearrange and synthesize, then draw a new sample and add it back to the data set. SMOTE (T, N, k) table input variables of this algorithm contain a small number of samples T, increase N%, and when synthesizing samples K nearest neighbor samples to refer to.

For example, as figure shown below, the left side is the original data; the darker individuals in the middle image are used to find positive individuals for k-nearest neighbors, here suppose we choose k=3, SMOTE algorithm will first identify the three nearest neighbors, next, one of the neighbors will be randomly selected to generate a new sample, finally, a new individual will be randomly generated on the line between the picked individual and the corresponding neighbor, then treat this individual as positive; when we select many different individuals to find k-nearest neighbors and synthesize new samples,

the final result will be shown as Figure 1.

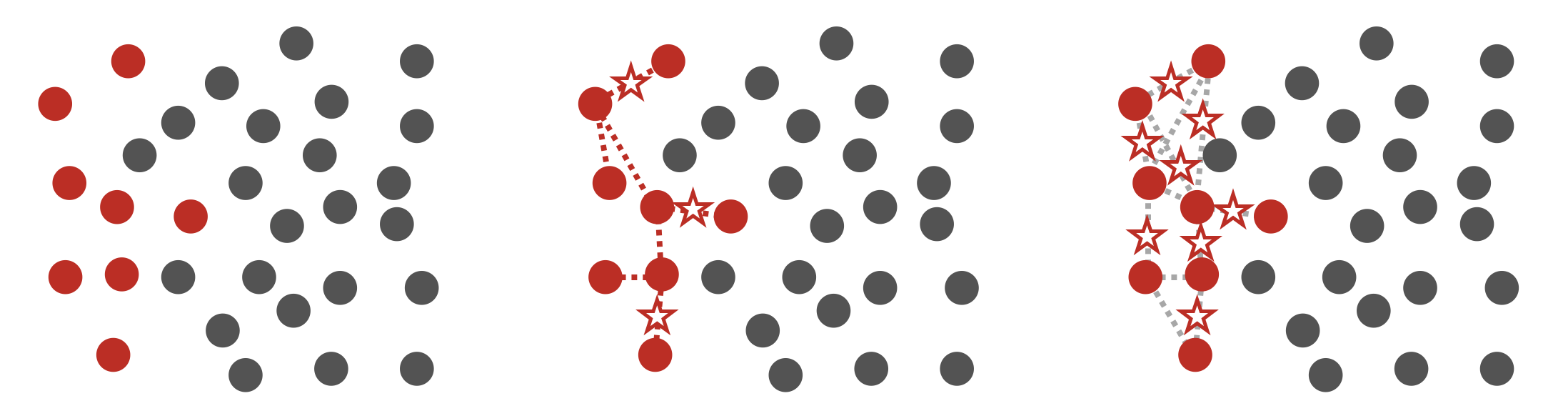


Figure 1 SVM example

El-Sayed et al. (2015) used SMOTE for sample augmentation in the study for processing autism imbalance data, and improve accuracy and credibility, the results show that the use of SMOTE is reliable. Dong (2015) uses DNA features to predict prostate cancer, the original data is high-dimensional and unbalanced in category.

Sarakit et al. (2015) classify text sentiment on YouTube, due to imbalance in the data set, so use SMOTE to balance, and use three machine learning classifiers, which is Multi- Naive Bayes, decision tree and support vector machine for testing, the results show that SMOTE technology can solve the problem of unbalanced data and obtain effectively improved classification effects. Based on the above, this study uses SMOTE for category balance.

2.2 Naive Bayes and Bayes Net

The core of this classification algorithm is based on Bayes' theorem. It needs to assume the prior probability of a known target category, this probability is often known from training samples. Under any given target category, the attributes of their participation are assumed to be independent of each other, this point echoes the assumption of Bayes' theorem. Because of this method’s construction is fast, so it is suitable for high-dimensional data classification. From the formula P(C | X) = P(X | C)\*P(C)/P(X), suppose there is a set of X attributes in the training data = {X1, X2, ..., Xk}, X does not contain the target category attribute, and C is the set of attribute values ​​of the target category, C={C1, C2, ..., Cn}, for this study, C contains two types of casualties and no casualties. P(C | X) is a set of X attributes, probability of target category C. When forecasting data,

after comparing the attributes {C1,C2,...,Cn} of each target category in the C set, selecting the attribute value of the target category with the highest probability as the prediction classification result. P(X | C) is probability of occurrence of an X attribute set in certain target category C. P(X) is the probability of cases appearing in a certain X attribute set.

Bayes Net is a probability graph based on a directed non-cycle relationship. The difference between Naive Bayes and Bayes Net is that there is a mutual influence and non-cycle between attributes, this relationship must be known in advance. For example, Bayes Net can be used to express the probability relationship between disease and its related symptoms;

if a certain symptom is known, the Bayes Net can be used to calculate the probability of various diseases. Due to the dependency relationship between attributes, so compared to Naive Bayes, more able to highlight the target hypothesis that the decision maker is interested in.

2.3 Support Vector machine

Support Vector Machine (SVM) is a kind of spatial transformation of high feature dimensions and correspond to a hyperplane to facilitate the supervised learning method of classification actions. //The main operation mode is to separate different target category points on both sides of the hyperplane, and its kernel functions are linear, polynomial, radius type and tangent type, etc. Taking the linear method as an example, suppose a hyperplane is f(x): W．X + b = 0, X is a high-dimensional coordinate, W is a vector, and b is a constant. The hyperplane f(x) can separate different sets of high-dimensional categories. This cutting choice needs to be the easiest to cut out the two-party categories as the priority. In simple terms, finding a decision boundary to make the boundary between the two categories (margins) maximized so that they can be perfectly separated. In this formula, the minimum values ​​of W and b are often solved by the Lagrangian method. In formula (1) is the judgment formula obtained by solving using Lagrange method L(a) and combining the classification attributes yi and yj. In this formula, ai and aj are Lagrangian variables, this variable needs to be assumed to have been obtained, xi and xj are high-dimensional coordinates, n is the total number of points corresponding to this hyperplane (the positive category is set to 1, the negative category is set to -1), and k(xi, xj) is the kernel function formula of SVM. The variables  and r are the core parameters, T is the transposed matrix, and d is the number of attribute dimensions.

|  |  |
| --- | --- |
|  | (1) |

2.4 C4.5, LMT and Random Decision Tree

C4.5 is an algorithm for building decision trees (Quinlan, 1993), its predecessor is the ID3 algorithm. The reason for using C4.5 as the classification calculation method in this study. In addition to the classification method based on ID3, which is often used, the classification model established by C4.5 is rarely discussed in the study of domestic aviation accident factors.

First of all, since the C4.5 algorithm will involve the basis of the ID3 algorithm, first, briefly describe the basis of the ID3 algorithm. The strategy of selecting attribute nodes adopted by the ID3 algorithm is information profit. The formula for information profit is shown in the following defined as , when information profit is substituted into a certain test attribute A, this formula subtracts the information entropy before the attribute A is divided and the information entropy after the division to obtain the Gain value. Information entropy before attribute A segmentation in si is the number of times under the i-th classification value, assuming that S is the sum of the times of {s1,..., sm}, Pi=si/S. Information entropy after attribute A segmentation in (2), it is similar to the variable of information entropy before segmentation, but attribute A should be used as the main cutting function, and the n value ranges of attribute A should be calculated separately. Where j is the serial number of an attribute value in n value ranges, i is the serial number of the target category of an attribute A, and m is the total number of classification categories. sij is the number of times that an attribute A appears in the i-th target category corresponding to the j-th attribute value.

|  |  |
| --- | --- |
|  | (2) |
|  |

The C4.5 algorithm is based on the improved ID3 algorithm. The method of selecting attributes is to use the Gain Ratio = Gain(A) / SplitInfo(A). The information used is profitable, and is combined with SplitInfo (3) to obtain this value. In the split information formula, there are n kinds of value ranges under the A attribute, S is the total number of data, and Sj is the number of occurrences of the jth type in the value range subset under the A attribute.

|  |  |
| --- | --- |
|  | (3) |

Logistic model trees are a classification model in computer science, which combines logistic regression and decision tree algorithm. The logical model tree is based on the early ideas of the model tree, and a linear regression model is provided on the leaves of the decision number to provide a piecewise linear regression model. The basic logical model tree induction algorithm uses cross-validation to find many LogitBoost iterations that do not overfit the training data.

In machine learning, Random Decision Tree is a classifier that contains multiple decision trees, and its output category result is determined by the mode of the output category of each tree. The advantages of random forest are as follows:

* Can handle multi-dimensional input variables.
* When the category is determined, you can assess the importance of the variable.
* It can still maintain accuracy even when most of the data is lost.
* For unbalanced classification sets, it can balance errors.
* It is very useful for data exploration, detecting outliers and visualizing data.
* The learning process is fast.

2.5 Radial Basis Function Network and Multilayer Perceptron

Neural network is a kind of classification method that simulates biological nerve cells. Its main structure includes input layer, hidden layer and output layer. These layers are connected to each other by a node and formed unidirectionally, and each node is connected by a connecting line, and these connecting lines are weighted. The input data is processed in the hidden layer before it can be classified in the output layer. The functions used in the hidden layer are represented as , m is the total number of classification categories, n is the total number of input values, Yj is the output target classification attribute value, function f converts the value calculated in the function into the classification category representative symbol, Wij is the input value weight, Xi is the input attribute value , Өj is the offset. In recent years, neural-like networks have been regarded as very effective non-linear model building tools. Therefore, this study uses a radial basis function network, the formula is shown below represented as , m is the total number of classification categories, K is the total number of hidden layer nodes, function F is the conversion function, Wkj is the weight value, which is the weight addition for the connection from the hidden layer to the output layer, and φk (x,ck) is the radial basis Function, x is the input value, ck is the center point of the k-th node of the hidden layer, this center point function is used for solving with the cluster method, Өj is the offset.

Multilayer Perceptron (MLP) is an artificial neural network. A multi-layer perceptron can be treated as a directed graph, composed of multiple node layers, each layer is fully connected to the next layer. Except for the input node, there is a neuron with a nonlinear activation function in each node. The multilayer perceptron itself can use any form of activation function, but in order to use the reverse conduction algorithm for effective learning, the activation function must be limited to a differentiable function.

MLP was a very popular machine learning method in the past, because it has a wide range of applications. However, in the later 90s, the emergence of support vector machines made multilayer perceptrons encounter A strong opponent. Until recently, due to the success of deep learning, multilayer perceptrons have regained attention. Because of this, this experiment also incorporates the multilayer perceptron into the experimental method.

**III. experimental design**

3.1 Experiment process

Step 1: Data collection

The research data was taken from the Accident/Incident Data System (AIDS) provided by the Federal Aviation Administration (2016). Only common Boeing flight accidents are considered. The period is from January 2000 to December 2019, with a total of 1209 transactions.

Step 2: Data preprocessing

In this study, various algorithms are executed using Weka and compared with each other. Weka provides several input formats, such as arff or csv, and the format used in this study is csv. After confirming the input format, you need to delete symbols that cannot be imported into Weka (eg: %, ‘) If the missing value in the field is too serious, delete this field. The number of other missing values ​​is also considered to be reduced. A small part of the missing value is replaced by UNKOWN attribute value in the field. Then execute SMOTE to make the ratio of casualties the same as the number of non-casualities, and then solve the problem of data imbalance. At this time, the total number of data is 1463. After the data is balanced, start to delete the field. First, it will be like the identification number of a single flight accident event or the registration number of the flight vehicle. Such intuitions will not directly or indirectly affect flight safety accidents. We will delete the affected fields. Then, use Information Gain to assess the importance of attributes, reduce the dimension of the data, and delete the fields with lower importance. There are 27 kinds of original data attributes, and 26 kinds after combining casualty 2 attributes. At this stage, invalid attributes (information profit is 0) and non-important attributes (such as the identification code of a single flight accident) will be removed.

Step 3: Use classification algorithms to model actions

Bayesian network, simple Bayesian, support for SVM, decision tree (C4.5), radial basis function network (RBFN) and multilayer perceptron (Multilayer Perceptron) to establish classification models, The evaluation index (overall classification accuracy, Recall, F-measure, AUC) is used to evaluate the effectiveness of this model. A 10-fold cross validation method is used for sample training and testing.

Step 4: Result discussion

The final stage will explain the experimental results. In this study, we look forward to comparing each other, discovering the different characteristics obtained by each method and further discussing the selected attributes.

3.2 Evaluate performance indicators

This research indicator is F-measure (4), which combines Precision (5) and Recall (6). The F-measure value range is 0 to 1. The higher the score, the better. AUC is the area under the ROC line. A higher score indicates that the ROC tends to the upper left, that is, the better the classification efficiency. In the end, we will use the Apriori correlation rule to understand the more frequent association combinations in the existing Flight Safety accidents. Perhaps we can find some common features that cause Flight safety accident and further speculate about the potential problems of each feature in the combination.

|  |  |
| --- | --- |
|  | (4) |

|  |  |
| --- | --- |
|  | (5) |

|  |  |
| --- | --- |
|  | (6) |

**IV. Experimental Results**

Table 1

Classifiers results comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Precision | Recall | F-measure | AUC | Accuracy |
| Bayes Network | 0.952 | 0.951 | 0.951 | 0.987 | 95% |
| Hidden Naive Bayes | 0.963 | 0.962 | 0.962 | 0.990 | 96.24% |
| Naive Bayes | 0.949 | 0.947 | 0.947 | 0.987 | 94.73% |
| RBF Network | 0.953 | 0.952 | 0.952 | 0.975 | 95.21% |
| SMO | 0.963 | 0.963 | 0.963 | 0.963 | 96.31% |
| J48 | 0.965 | 0.965 | 0.965 | 0.977 | 96.51% |
| LMT | 0.961 | 0.961 | 0.961 | 0.980 | 96.1% |
| Random Forest | 0.964 | 0.964 | 0.964 | 0.988 | 96.45% |
| Random Tree | 0.951 | 0.950 | 0.950 | 0.976 | 95% |
| Multilayer Perceptron | 0.954 | 0.954 | 0.954 | 0.985 | 95.35% |

Table 2

Apriori association results

|  |  |  |
| --- | --- | --- |
| Input attributes | | Output |
| Primary Flight Type=SCHEDULED AIR CARRIER 30 | | Is\_Fatal = 1 |
| Primary Flight Type=SCHEDULED AIR CARRIER | Nbr of Engines=2 29 |
| Aircraft Engine Make=RROYCE 24 |
| Aircraft Engine Make=RROYCE 24 | |
| Aircraft Engine Model=UNKNOW 24 | |

**V. Conclusion**

After screening by Information Gain and some people, this study found that "airport where the event occurred", "city where the event occurred" and "state where the event occurred", these field characteristics have a greater impact on Flight Safety. Area-related attributes such as city, state and airport where the event occurred may be related to the density of routes in the area.

Another part of the indirect human relationship of attributes, such as the Operator here (the airline to which the flight vehicle is affiliated), may be related to the operator's organizational policy. Aircraft Model (vehicle model) may be related to the structural design of the model.

Before SMOTE was used, the overall data set of casualties accounted for very few, resulting in poor classification of the classification results and no suitable features. But after using SMOTE, the classification results are much better. Among the classifier comparison results of this experiment, we found that the classification result of J48 is the best. The type of flight, the number of engines, and the engine manufacturer are important factors leading to Flight Safety accident in the association rules.

The source of this experiment is the AIDS database, because the content is limited and no more attributes can be provided for exploration, so some potential attributes that cause Flight Safety accident casualties may still exist. About this part, I hope that more attributes will be provided in the future Exploration research.

**References**

IATA Safety Report，2015，52nd Edition.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research, 16*, 321-357.

El-Sayed, A. A., Mahmood, M. A. M., Meguid, N. A., & Hefny, H. A. (2015, November). Handling autism imbalanced data using synthetic minority over-sampling technique (SMOTE). In *2015 Third World Conference on Complex Systems (WCCS)* (pp. 1-5). IEEE.

Federal Aviation Administration, Accident and Incident Data System (AIDS),January 2016 (available online at http://www.asias.faa.gov/pls/apex/f?p=100:12:0::NO:::)

Dong, S. (2015). Exploring effects of various data preprocessing methods on the classification of DNA microarray data. In *Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP)* (p. 74). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

Salzberg, S. L. (1994). C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993.

Raorane, A. A., Kulkarni, R. V., & Jitkar, B. D. (2012). Association rule–extracting knowledge using market basket analysis. *Research Journal of Recent Sciences ISSN, 2277*, 2502.

Sarakit, P., Theeramunkong, T., & Haruechaiyasak, C. (2015, August). Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)* (pp. 1-5). IEEE.