# *ANALYSIS OF FEATURE SELECTION TECHNIQUES IN CREDIT RISK ASSESSMENT*

R.S.Ramya
PG scholar/Department of CSE
Government college of Technology, Coimbatore.
e-mail:ramyariya19@gmail.com

Professor S.Kumaresan
Head of the Department/Department of CSE
Government college of Technology, Coimbatore.
e-mail:sukumaresan@gct.ac.in

***Abstract*-Data Mining is an automated extraction of hidden knowledge from large amount of data. The computational complexity of the data mining algorithms increases rapidly as the number of features in the dataset increases. Real world credit datasets have accumulated large quantities of information about clients and their financial and payment history. Feature selection techniques are used on such high dimensional data to reduce the dimensionality by removing irrelevant and redundant features to improve the predictive accuracy of data mining algorithms. The objective of this work is study the information gain, gain ratio and chi square correlation based feature selection method to reduce the feature dimensionality.**

**Information gain measure identifies the entropy value of each specific feature. The amount of information gain or entropy is used to decide whether the feature is selected or deleted. Gain ratio applies normalization technique to information gain using spilt information value. The correlation based feature selection uses heuristic search strategies to estimate how the features are correlated with the class attribute and how they are important of each other. Experiments were conducted on the German credit dataset available at UCI Machine Learning Repository to reduce the feature dimensionality using these feature selection methods.**

***Keywords—Data Mining, Credit risk assessment, Feature selection, Information gain, Gain ratio, Chi square correlation.***

## I. INTRODUCTION

The tremendous growth in computing power and storage capacity has resulted in the growth of huge databases**.** Data Mining, also popularly referred to as Knowledge Discovery from Data (KDD), is the automated extraction of patterns from large amount of data. Data mining methods are used in different areas such as business, data management, scientific, engineering, banks, government administration and many other applications. The KDD process consists of an iterative sequence of the steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge representation. Data mining is an essential step in the process of knowledge discovery in data bases, in which intelligent techniques are applied in order to extract useful patterns.

In general, data mining tasks can be grouped in two categories: descriptive and predictive. Descriptive mining tasks include characterized general properties of the data. Examples of descriptive mining tasks include association mining and cluster analysis. The inference from the current data is used by the predictive task in order to make prediction. Examples of predictive mining task include classification and prediction.

In many practical situations there are far too many features for learning algorithms to handle, some of them are irrelevant or redundant for mining. Hence, it is necessary to select a subset of features which are relevant to use in learning. The performance, robustness and usefulness of data mining algorithms are improved when relatively few and relevant features are involved in the process. Literature shows that irrelevant or redundant features cause the performance of learning algorithms to deteriorate.

## II. FEATURE SELECTION

Feature selection also known as feature reduction is a technique of selecting a subset of relevant features from the original dataset. Feature selection process removes the irrelevant and redundant features and increases the predictive accuracy of the data mining algorithm.

To evaluate the quality of a subset of features, the feature selection methods mainly adopt the following approaches: the filter and the wrapper. In filter method, the feature set is filtered to select the most informative subset before learning commences based on general characteristics of the data itself and independent of the mining algorithm. In wrapper method, the learning algorithm is wrapped into the selection procedure it uses the performance mining algorithm to evaluate and determine the quality of features.

Filter methods evaluate feature individually based its correlation with the target attribute and then select features with highest value. They are generally faster and more practical to use on high dimensional datasets. In this study we use information gain, gain ratio and correlation based feature selection measures of filter methods.

## III. METHODOLOGY

The original dataset is given as a input to feature selection technique such as information gain, gain ratio, chi square

[978-1-4799-6438-3/15/$31.00 ©2015 IEEE]

correlation. Three reduced feature subsets are obtained from the above three feature selection methods.

The model work flow diagram is described in figure 1

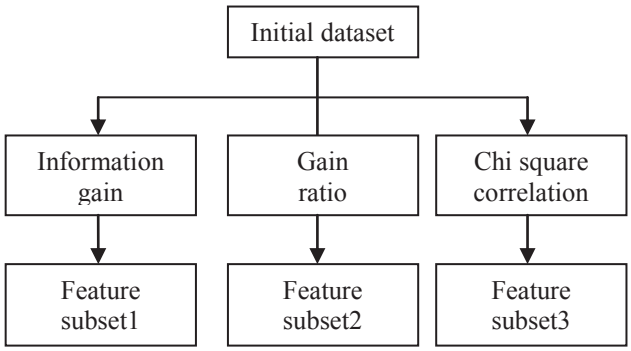


Figure 1 Model Flow Diagram

## IV. DATASET DESRIPTION

The German credit dataset is a public domain dataset available in the well known data repository of the UCI (University of California and Irvine). It consists of 1000 instances where each instance contains details about the bank clients. The dataset is composed of 700 instances of creditworthy applicants and 300 instances of bad credit applicants. The dataset contains 20 attributes including one class attribute. The class attribute indicates whether the bank client is credit worthy or not.

Table 1 German credit dataset description

| DATASET | GERMAN CREDIT |
|---|---|
| Attribute type | Categorical, Numerical |
| Number of Attribute | 20 |
| Number of instances | 1000 |
| Missing values | No |
| Number of classes | 2 |

The non-class attributes are Status of existing checking account, Duration in month, Credit history, Purpose, Credit amount, Savings account, Present employment since, Installment rate in percentage of disposable income, Personal status and sex, Other debtors / guarantors, Present residence since, Property, Age, Other installment plans, Housing, Number of existing credits at this bank, Number of people being liable to provide maintenance for, Telephone, Foreign worker. The description of each attribute of the dataset is given in Table 2

Table 2 German credit dataset attribute description

| S.No | Name of the attribute | Number of distinct values | Type of the attribute |
|---|---|---|---|
| 1. | Status of checking account | 4 | Categorical |
| 2. | Duration in month | 33 | Numerical |
| 3. | Credit history | 5 | Categorical |
| 4. | Purpose | 10 | Categorical |
| 5. | Credit amount | 921 | Numerical |
| 6. | Savings account | 5 | Categorical |
| 7. | Present employment since | 5 | Categorical |
| 8. | Installment rate in percentage | 4 | Numerical |
| 9. | Personal status and sex | 4 | Categorical |
| 10. | Others debtors | 3 | Categorical |
| 11. | Present residence since | 4 | Numerical |
| 12. | Property | 4 | Categorical |
| 13. | Age in year | 53 | Numerical |
| 14. | Other installment plans | 3 | Categorical |
| 15. | Housing | 3 | Categorical |
| 16. | Number of existing credits at this bank | 4 | Numerical |
| 17. | Job | 4 | Categorical |
| 18. | Number of people being liable to provide maintenance | 2 | Numerical |
| 19. | Telephone | 2 | Categorical |
| 20. | Foreign worker | 2 | Categorical |
| 21. | Class | 2 | Numerical |

## V. INFORMATION GAIN

Information gain is a feature selection technique which ranks each feature based on the given training tuples. Information gain is based on Claude Shannon's work on information theory, which studied the value or information content of tuples. The information needed to classify the tuples into partitions using the feature with highest information gain is minimum and the resulting partitions will also contain lowest degree impurity or randomness. This approach minimizes the expected number of tests needed to be performed in the classification process.

The expected information needed to classify a tuple in D is given by

$$Info(D) = - \sum_{i=1}^{m} pi \log_2(pi) \quad -(1)$$

The probability that a particular tuple in D belongs to class ci is estimated using the ratio $|ci,D|/|D|$ and is represented as pi. The log function with base 2 is used, because the information is encoded in bits. Info(D) indicates the average information required to identify the class of a tuple in database. The information is based on the proportions of tuples of each class. Info(D) is also known as the entropy of D.

The tuples of D is partitioned on some feature A having v distinct values {a1,a2,.....av}. The amount of information required to arrive at the exact classification after this partitioning is measured using the formula 2

$$InfoA(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} * Info(Dj) \quad -(2)$$

The term $|Dj|/|D|$ represent the weight of the partition. The expected amount of information required to classify a tuple based on the partitioning on the attibute A from database D is represented as InfoA(D). The smaller the required amount of expected information, the greater is the purity of the partitions. Information gain is defined as the difference between the original requirement of information (based on the proportion of classes) and the new information requirement (obtained after partitioning on A)

$$Gain(A) = Info(D) - InfoA(D) \quad -(3)$$

Gain(A) indicates how much information would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A. The feature A with the higher information gain (Gain(A) ) is better ranked at the training tuples.

To calculate the entropy of a attribute, we have to first determine the number of different instances in a attribute and number of tuples belonging to each of the instances of the attribute. The number of tuples associated with a particular instance of a attribute is further divided based on class to which the tuple belongs.

The value of original entropy (Info(D)) can be calculated by determining the number of tuples of class 1 and number of tuples of class 2. In the German credit dataset of the 1000 instances 700 tuples belong to class 1 and 300 tuples belong to class 2. Table 3 gives details about the number of distinct instances in the Status of existing checking account attribute(A1) and also the number of tuples belonging to class 1 and class 2 for each distinct instance of that attribute.

Table 3 Description of Status of existing checking account attribute

| Number of distinct instances | Number of tuples | Number of tuples of class 1 | Number of tuples of class 2 |
|---|---|---|---|
| A11 | 274 | 139 | 135 |
| A12 | 269 | 164 | 105 |
| A13 | 63 | 49 | 14 |
| A14 | 394 | 348 | 46 |

The details from the table above and the calculated original entropy value (Info(D)) is used to calculate the expected new entropy of the Status of existing account attribute(InfoA1( D)). The difference between original entropy and the expected new entropy gives the information gain of the attribute. This process is repeated for all the non-class attributes of the German credit dataset. Table 4 lists the attributes of the German credit dataset in descending order of the value of the calculated information gain.

Table 4 Information gain ranking.

| **Attribute No** | **Name of the attribute** | **Value of information gain** |
|---|---|---|
| 1 | Status of checking account | 0.094739 |
| 3 | Credit history | 0.043618 |
| 2 | Duration in month | 0.0329 |
| 6 | Savings account | 0.028115 |
| 4 | Purpose | 0.024894 |
| 5 | Credit amount | 0.018709 |
| 12 | Property | 0.016985 |
| 7 | Present employment since | 0.013102 |
| 15 | Housing | 0.012753 |
| 13 | Age in year | 0.011278 |
| 14 | Other installment plans | 0.008875 |
| 9 | Personal status and sex | 0.006811 |
| 20 | Foreign worker | 0.005823 |
| 10 | Others debtors | 0.004797 |
| 17 | Job | 0.001337 |
| 19 | Telephone | 0.000964 |
| 18 | Number of people being liable to provide maintenance | 0 |
| 8 | Installment rate in percentage | 0 |
| 11 | Present residence since | 0 |
| 16 | Number of existing credits at this bank | 0 |

The top 12 attributes of high information gain are selected as a reduced feature subset.

## VI. GAIN RATIO

The Information gain technique is biased towards test with many outcomes. An extension to information gain is known as gain ratio which attempts to overcome this bias nature. It normalizes the information gain using a "split information" value defined based on Info(D) using formula 4

$$\text{SplitInfoA(D)} -= \sum_{j=1}^{v} \frac{|Dj|}{|D|} * \log_2 \frac{|Dj|}{|D|} \quad \text{-(4)}$$

This value considers the number of tuples having that outcome with respect to the total number of tuples in D. The term $|Dj|/|D|$ acts as the weight of the jth partition. The gain ratio of a attribute is defined as the ratio between information gain and the split information of that attribute. It is calculated using formula 5

$$\text{GainRatio(A)} = \text{Gain(A)} / \text{SplitInfo(A)} \quad \text{-(5)}$$

where Gain(A) is the information gain of the attribute A. The feature with the higher gain ratio is better ranked at the given training tuples.

To calculate the gainratio of a attribute, we have to first calculate the information gain value (Gain(A)) of the attributes of the German credit dataset. The procedure to calculate information gain value has been explained in detail in chapter1. Then determine the number of different instances in a attribute and number of tuples belonging to each of the instances of the attribute and using these details the splitinfoA(D) value of the attribute is calculated. The ratio of splitinfoA(D) and Gain(A) values gives the final value of the gain ratio.

Table 3 gives details about the number of distinct instances in the attribute Status of existing checking account (A1) and also the number the tuples belonging to class 1 and class 2 for each distinct instance of that attribute. The details from the table 3 and the used to calculate splitinfoA(D) of the Status of existing account attribute. Table 5 lists the attributes of the German credit dataset in descending order of the value of the calculated gain ratio. The top 12 attributes of high gain ratio are selected as a reduced feature subset.

## VII. CHI SQUARE CORRELATION

Correlation-based feature selection is a filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function. It does not require the user to specify any thresholds or the number of features selected. It uses heuristic search strategies to search the best combination of features.

Table 5 Gain ratio ranking

| Attribute No | Name of the attribute | Value of gain ratio |
|---|---|---|
| 1 | Status of checking account | 0.052573 |
| 20 | Foreign worker | 0.025499 |
| 3 | Credit history | 0.02548 |
| 2 | Duration in month | 0.025458 |
| 5 | Credit amount | 0.022629 |
| 6 | Savings account | 0.016658 |
| 13 | Age in years | 0.016078 |
| 15 | Housing | 0.011197 |
| 14 | Other installment plans | 0.010507 |
| 4 | Purpose | 0.009335 |
| 10 | Others debtors | 0.008909 |
| 12 | Property | 0.00872 |
| 7 | Present employment since | 0.006079 |
| 9 | Personal status and sex | 0.004445 |
| 19 | Telephone | 0.00099 |
| 17 | Job | 0.000946 |
| 16 | Number of existing credits at bank | 0 |
| 8 | Installment rate in percentage | 0 |
| 11 | Present residence since | 0 |
| 18 | Number of people being liable to provide maintenance | 0 |

The chi-square statistics is a correlation method used to show whether or not there is a relationship between two attributes in a dataset. The chi square correlation measures the deviation of observed frequency distribution from the theoretical distribution which is represented using the formula

$$\chi^2 = \sum_{i=1}^{n} \frac{(Oi - Ei)^2}{Ei} \quad \text{-(6)}$$

where $\chi 2$ gives the chi square correlation value, Oi indicates the observed frequency, Ei indicates the expected frequency. The value of Ei is calculated using the values of row total and column total using the formula 7

Ei = Row total * Column total / (Total of all the entries) -(7)

The attribute with the higher chi square correlation value is better ranked at the given training tuples than the attributes of the dataset.

To calculate the chi square correlation value of a attribute, we have to first determine the number of different instances in a attribute and number of tuples belonging to each of the instances of the attribute. The number of tuples associated with a particular instance of a attribute is further divided based on class to which the tuple belongs. These details are used to represent the original frequency value in a matrix form where the rows represents the target attribute class instances and the column represents the different instances of a non class attribute.

Using this matrix the expected value is calculated using the formula and represented in matrix form similar to that of the original frequency representation. Then the original frequency and the expected frequency values are used to calculate chi square correlation as mentioned in the formula.

Table 7 gives details about the number of distinct instances in the attribute Foreign worker (A20) and also the number the tuples belonging to class 1 and class 2 for each distinct instance of that attribute. The details from the table and the used to calculate chi square correlation value of the Foreign worker attribute.

Table 7 Description of Foreign worker attribute

| Number of distinct instances | Number of tuples | Number of tuples of class 1 | Number of tuples of class 2 |
|---|---|---|---|
| A201 | 963 | 667 | 296 |
| A201 | 37 | 33 | 4 |

Table 8 lists the attributes of the German credit dataset in descending order of the value of the calculated chi square correlation. The top 12 attributes of high chi square correlation value are selected as a reduced feature subset.

## VIII. CONCLUSION AND FUTURE WORK

The feature selection techniques such as information gain, gain ratio, chi square correlation were applied to the German credit dataset available in UCI Machine Learning Repository. These feature selection techniques selected the features that will be useful for classification of clients and the ones that are irrelevant and redundant are omitted. The three feature subsets obtained by using the above feature selection techniques are compared with each other and is concluded that the three feature selection techniques produces almost the same feature subset. The union of the three subsets is obtained.

As future work the resulting reduced feature subset can be further refined using heuristics based feature selection technique such as genetic algorithm and the final refined feature subset can be obtained. This final feature subset can used as a input for classification methods such as Artificial neural networks, Decision tree, Support vector machine. The accuracy of classification by these different classifiers on reduced feature subset can be analyzed.

Table 8 chi square correlation ranking

| **Attribute No** | **Name of the attribute** | **Value of gain ratio** |
|---|---|---|
| 1 | Status of checking account | 123.7209 |
| 3 | Credit history | 61.6914 |
| 2 | Duration in month | 46.8311 |
| 6 | Savings account | 36.0989 |
| 4 | Purpose | 33.3564 |
| 5 | Credit amount | 26.9528 |
| 12 | Property | 23.7196 |
| 7 | Present employment since | 18.3683 |
| 15 | Housing | 18.1998 |
| 13 | Age in year | 16.3681 |
| 14 | Other installment plans | 12.8392 |
| 9 | Personal status and sex | 9.6052 |
| 20 | Foreign worker | 6.737 |
| 10 | Others debtors | 6.6454 |
| 17 | Job | 1.8852 |
| 19 | Telephone | 1.3298 |
| 18 | Number of people being liable to provide maintenance | 0 |
| 8 | Installment rate in percentage | 0 |
| 11 | Present residence since | 0 |
| 16 | Number of existing credits at this bank | 0 |

## IX. REFERENCES


[1] Stjepan Oreski, Goran Oreski "Genetic algorithm-based heuristic for feature selection in credit risk assessment", Expert Systems with Application, vol:41,(pp 2052–2064),Elsevier,2014.

[2] Stjepan Oreski, Goran Oreski "Hybrid system with genetic algorithm and arificial neural networks and its application to retail credit risk assessment", Expert Systems with Application, vol:39,(pp 12605–12617),Elsevier,2012.

[3] K.C Tan, E.J. Teoh, Q.Yu, K.C.Goh "A hybrid evolutionary algorithm for attribute selection in data mining" Expert Systems with Applications, 36, 8616-8630, 2009.

[4] Ekrem Duman, M.Hamdi Ozecelik "Detecting credit card fraud by genetic algorithm and scatter search" Expert Systems with Applications,38, 13057-13063,2011.

[5] Pappa.G.L, Freitas A.A, Kaestner C.A.A "A multi objective genetic algorithm for attribute selection" In proceedings of the fourth International conference on recent advances in soft computing (RASC-2002)(pp.116-121) Berlin: Springer,2002.

[6] I.Cheng Yeh, Che-hui Lien "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients" Expert Systems with Applications, 36, 2473-2480,2009.

[7] Lee T.S, Chiu C.C, Lu C.J, Chen I.F "Credit scoring using the hybrid neural discriminant technique" Expert Systems with Applications, 23(3), 245-254,2002.

[8] Adnan Khashman "Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes" Expert Systems with Applications, 37, 6233-6239,2010.

[9] Chia-Ming Wang, Yin-Fu Huang "Evolutionary based feature selection approaches with new criteria for data mining: A case study of credit approval data" Expert Systems with Applications, 36, 5900-5908,2009.

[10] Cheng-Lung Huang, Chieh-Jen Wang " A GA-based feature selection and parameter optimization for support vector machines" Expert Systems with Applications, 31, 231-240,2006.

[11] Hua Tang, Jun Lu " A hybrid algorithm combined Genetic algorithm with Information Entropy for Data mining" Second IEEE conference on Industrial Electronics and Applications, 753-757,2007.

[12] Yu.L, Liu H"Efficient feature selection via analysis of relevance and redundancy" Journal of Machine Learning Research, 5, 1205-1224,2004.

[13] Frohlich H, Chapelle O "Feature selection for support vector machines by means of Genetic Algorithms" In Proceedings of the fifteenth IEEE International Conference on Tools with Artificial Intelligence, USA, 142-148,2003.

[14] Guyon I, Elisseeff A " An Introduction of variable and feature selection" Journal of Machine learning Research" vol 3, 1157-1182,2003.

[15] Han J, Kamber M "Data Mining: Concepts and Techniques" Morgan Kaufman,2010.

[16] Mannila H "Theoritical Frameworks for Data mining SIGKDD Explorations Journal" vol 1,issue 2, 30-32,2002.

[17] Yu L, Wang S.Y, Lai K.K "Credit risk assessment with a multistage neural network ensemble learning approach" Expert Systems with Applications, 34(2), 1434 -1444,2008.

[18] Tsai C.F, Wu J.W "Using Neural network ensembles for bankruptcy prediction and credit scoring" Expert Systems with Applications, 34(4), 2639-2649,2008.

[19] Xu X, Zlou C, Wang Z "Credit scoring algorithm based on link analysis ranking with support vector machines" Expert Systems with Applications, 36(2), 2625-2632,2009.

[20] Chuang C.L , Lin R.H "Constructing a reassigning credit scoring model" Expert Systems with Applications" 36(2), 1685-1694,2009.