An Approach for Efficient Knowledge Mining with the application PRM

[1] Neeraj Bhargava, [2] Dr. Ritu Bhargava, [3] Abhishek Kumar, [4] Prerna Pareek
[1] Professor, [2] Lecturer r, [3] Assistant Professor, [4] Research scholar
[1] School of Engineering & System Sciences, [2] Computer Science,
[1] MDS, University Ajmer, India, [2] Sohpia Girls' College Ajmer, India , [3] Aryabhatta College of Engg. & Research Center, Ajmer India, [4] MJRP University Jaipur

Abstract : The methodology of data mining is used for knowledgeable dredging of information in bulk amount. This became common in approach nowadays with all associated tools in order to obtain meaningful data from heterogeneous data set for example data warehouse, data mart, and repositories. The data is mined with an approach to better understand and analyze it. The operations which are applicable to the huge datasets are the different process of classification, associations, and other tools as well. In this paper, we basically focus on analyzing data based on the classification of predictive association rules, (CPAR). It is completely rule-based learning so in order to get better accuracy in results we apply various association based methods with classification techniques to get better results. This paper has result based on applying FOIL algorithm with the application of CPAR. The paper has primarily focused on the concept of decreasing complexity on huge dataset .the concept of Foil (First-order inductive learning) is inherited by CPAR which is powerful rule-based learning technique.

***Keywords* - FOIL, Data mining, CPAR, Association Rule.**

I. INTRODUCTION

FOIL (First-order inductive learner) was contemplated by Quinlan in 1993 which was somehow based on the greedy algorithm approach. Foil algorithm is working in such a way that it distinguishes positive examples from negative ones. The optimum part of the results searched by this particular algorithm and all positive rules are eliminated till all the positives rules of the data sets are covered completely.

Data mining is the process of extracting the hidden predictive information in such manner that the information act as knowledge. It is the innovative technology which is considered to be the great potential for any organization to deal with only relevant information by eliminating the useless data. It is very helpful for predicting the future trends, behaviors, and provides better support to the businesses activities in more optimum manner. The variation in the different types of algorithms in data mining it is a result of better research applications and product developments. There is the much profitable application of Data mining and many algorithms associated with it. This paper deals with the association rules applied over the dataset with the help of FOIL and CPAR algorithm. Association rule mining is a method which primarily aims to find frequent patterns correlations and frequent items sets associations; it provides an implicit structure of datasets available in different relational databases, and different data repositories available. So basically association rules are the creation of analysis of data for different frequency patterns, using the concept of support and confidence. [1]This particular rule is very much applicable and helpful in analyzing and predicting the probability of events to get happened. In this paper, the work primarily focuses the prediction of diseases symptoms and its consequences along with symptoms using FOIL method, in order to get the better result with accuracy and informative knowledge of available datasets. In this paper, we center around three essential issues in information mining including unverifiable information.

Recently, health information mining has picked up its essentialness. Dredging relevant facts and information which leads to accurate results out of bulk data

Utilizing knowledge learning to help in decision making for the conclusion, and provide the better pattern of data for efficient classifications. Association rule mining is a descriptive information mining assignment regularly utilized for market basket analysis. The depiction of the buy conduct of the client and the relationship among market– buyer things are revealed by the affiliation rules. A mining rule govern is a classification of the frame X->Y, where X and Y are thing sets, which are disjoint in nature. Characterization is a valuable information mining used to decide a display utilizing variable information to anticipate some results of classification. Association Classification (AC) is a current system, which coordinates the idea of arrangement and association which manage to mine. In this work, for health mining information mining, CPAR and Foil Algorithm development are the well-known association mining calculations used to remove Class predictive Association Rules (CPAR's).[2]The algorithm finds the connection between the diseases side effects that are helpful for the ailment. In spite of the fact that the association order is the effective classifier, it encounters less effective classifier exactness; since it creates the vast number of CPAR's to build the classifier. [3] In this paper, CPAR is proposed along with Foil to enhance the exactness of the current CPAR classifier for disease mining information analysis. In the meantime, deciding the nature of the guidelines is a critical assignment. It can be distinguished by utilizing proper mistake evaluate measures, for example, Laplace precision and Likelihood proportion measurements. In this work, Laplace precision is utilized for govern assessment. In light of Laplace precision, the best k-rules are chosen for classifier development. Dimensionality diminishment is a usually utilized pre-handling operation for any information mining errand. Lessening the number of properties by keeping the most critical ones dependably improves the nature of the results. Therefore, holding huge qualities in a dataset contributes high classifier precision. To accomplish that, factual test and calculation have been performed to recognize the critical properties of heart diseases datasets.[4] To finish up, the effect of both dimensionality diminishment methods with CPAR and CPAR has been broke down. From the different analysis it is discovered that the mix of FOIL and CPAR- yields better outcomes regarding classifier exactness.[5]

II. LITERATURE SURVEY

The previous works that has been done in past has been analyzed in this section with their strength and weaknesses .Different problem formulation has been depicted with the help of analysis in the next section.The working of FOIL algorithm and its associated algorithm are discussed in detail in order to get at some specific results. [9]The Foil classifier is applied for getting optimal solution over a given datasets

Authors proposed initially a framework in early nineties ,it was termed as associative classifications which was supposed to be integrations of association rules mining and classifications, The work depicted special subset of association rules in which the resultant part are restricted to the classifications class labels called class association rules. The proposed method states that when class label for samples are predicted the optimal rule is chosen.[10]

In 2009 Green B et.al has contemplated an algorithm which was classification based on multiple association s rules which used multiple class rules for associations for accurate and efficient classifications The method extent an efficient mining algorithm which was applicable for predicting unseen samples in an efficient manner.[11]

In 2010 G, et al. has been proposed a method for classification with predictive association rules, which has combined effects of both algorithm traditional rule and associative rule learning. Which proved to be advantageous in traditional learning rules? It was basically proposed in the work greedy algorithm was adopted by CPAR in order to generate directly from greedy data and then tends to test more rules which result in avoiding missing important rules, in this way optimum rules for prediction was evaluated in order to avoid overfitting. [12]

In 2015 Shen et.al yet another work proposed a new algorithm with commuted approach in traditional foil for the purpose of rules generation; it was suppose to provide much more efficient result than existing work. There were many rules with similar accuracy based on similar datasets from all iterations. [6]

Yuan et.al 2014 has proposed a combined solution for deleting the classified class and get the optimum solution with respect to the datasets .the authors proposed a solution for different class classification within the same data sets ,At the same time the classification accuracy is much better than the existing work comparatively [7].

Paladugu S (2010) has proposed a very different and nonconventional technique for getting better classifications of the different class of data sets compared to the average expected accuracy by using optimal k-rules of each class and the class was chosen with highest expected accuracy in form of predicted class. The paper has used best k rules of each class for prediction with different FOIL procedures applied to the given datasets. [8]

III. METHODOLOGY

This segment of Paper explained the Methodology part, in which the combinations are on the different methods that are applied over the dataset. These methods are required for the expectation of heart disease details in patients[13]

. Name	Type	Description

Age	Nominal	Min=20-40
		Avg=41-60
		Upper=61-80

Chest Pain	Nominal	Asympt atyp_angina non_anginal typ_angina

Rest b press	Numeric	Good=120/80 mm
		Average=129-139
		Poor=179/90

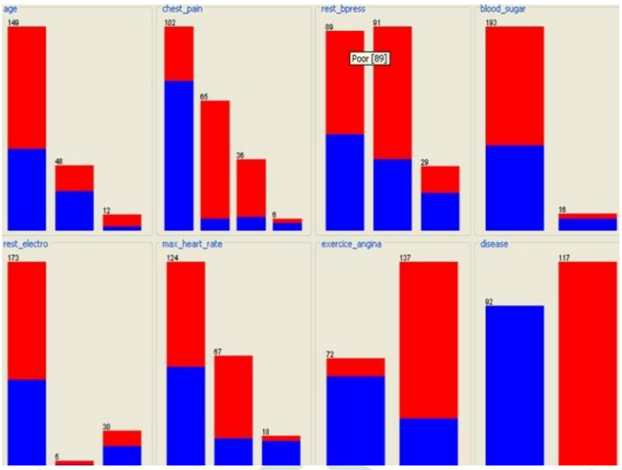
Blood Sugar rate	Nominal	F=False, T=True
Rest Electro	Nominal	Normal, left_vent_hyper, st_t_wave_abnormality
Maximum Heart Rate	Nominal	A<=100 B>100&<=150 C>150
Exercise Angina	Nominal	Yes, no
Disease	Nominal	Positive negative

The dataset of heart disease

There are eight fundamental characteristics of the dataset. Each trait with the exception of Disease is the fundamental driver of the Heart Disease. Each reason is ordered by some predefined measures. These measures are sorted by the making result proficient. [14]

1) Visualization of Heart Patients: Here, Heart Dataset is broke down outwardly utilizing distinctive properties and demonstrates the appropriation of qualities.

IV. RESULT



Results Analysis

The resulting analysis includes the patient data, the first step is entering the patient data into the tool, once the dataset is compatible with the tool then CPAR algorithm will be applied aptly in order to get the data classifications with better accuracy. The proposed methodology has calculated the probability of each attribute. CPAR has the ability to calculate the correlations between different target variables in very much efficient manner. The most unique attribute of CPAR algorithm it works much better in condition when the dimensionality of data items is high.[17]

1) Performance of the Classifier FOIL: In this particular analysis the duplicate rules generated are deleted in a manner that it removes useless and anomalies present in the databases. The hybrid application od CPAR and PRM many times can result in generating duplicate rules which is one of its drawbacks In that particular case the result is more dependent on the assumption we made.

Classifier	Evaluation Criteria	Value
FOPL	Timing to build the model (in Sec)	0.08
	Correctly Classified Instances	167
	Incorrectly Classified Instances	42
	Accuracy	64.1148%

V. PERFORMANCE OF CLASSIFIER FOIL

The above table containing the execution of the classifier, which incorporates some real execution measures like model building timing, accurately and erroneously examples and the primary significance component is exactness. This exactness component demonstrates the precision of grouping dataset. Since this is the outcome created by machine so the level of precision may change on another machine. This exactness or result or choice isn't as impeccable or correct as manual figuring. [15]

1) Estimates of FOIL: At first the optimal k rules are selected from each class of available datasets. These rules are applied over each class of data set in order to get better accuracy comparatively.

Here the characterization precision is appeared in table III, which depends on FOIL Classifier, is connected on the dataset, for example, In this segment of work Laplace of each rule is applied to get accuracy [18]

Evaluation Criteria	**Value**
Kappa Statistics	0.5913
Mean Absolute Error	0.2779
Root Mean Squared Error	0.3869
Relative Absolute Error	56.3624 %
Root Relative Squared Error	77.9345 %

The above table shows the evaluation criteria of classifier like kappa statistics, MAE, RMSE, RAE etc are calculated according to predefined formulas.[17]

Kappa Statistic:

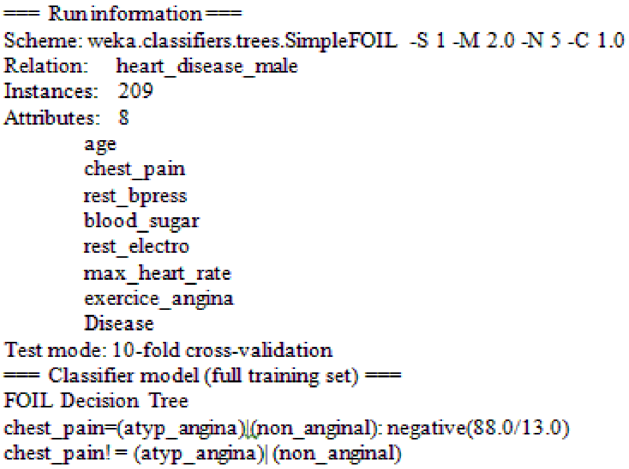
$$\boxed{K = \frac{P\,(A) - P\,(E)}{1 - P(E)}}$$

Root Mean Squared Error

$$\boxed{\frac{|p1-a1|+\ldots+|pn-an|}{n}}$$

Essential Accuracy Measures: Despite the simple applications of CPAR, this algorithm outperforms classifications methods many times for dynamic datasets also.This result is generated by a huge attributes datasets which is static in nature.[19]
The detailed accuracy measures by class using various factors like TP, FP, Precision, Recall and F-Measure.

Confusion Matrix: By using the classification results now confusion matrix is generated and calculated using different formulae of calculation correlations between different variables existing in the dataset.



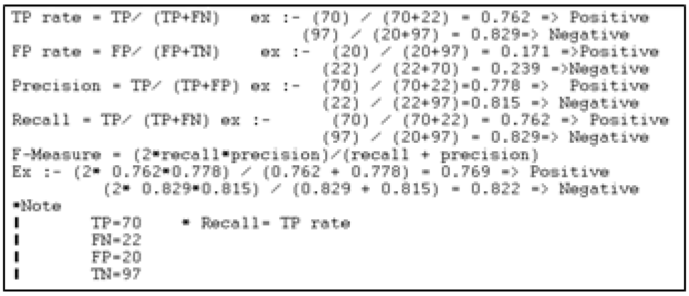
```
=== Run information ===
Scheme: weka.classifiers.trees.SimpleFOIL  -S 1 -M 2.0 -N 5 -C 1.0
Relation:    heart_disease_male
Instances:   209
Attributes:   8
             age
             chest_pain
             rest_bpress
             blood_sugar
             rest_electro
             max_heart_rate
             exercice_angina
             Disease
Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===
FOIL Decision Tree
chest_pain=(atyp_angina)|(non_anginal): negative(88.0/13.0)
chest_pain! = (atyp_angina)| (non_anginal)
```

The confusion matrix is achieved; calculate the accuracy, sensitivity and specificity measures. It classifies the accuracy of the model built for the prediction. Table Styles

Classifier	Positive	Negative	Class
FOIL	77	22	Positive
	20	97	Negative

Confusion matrix

Figure 2 Final results Calculation for FOIL

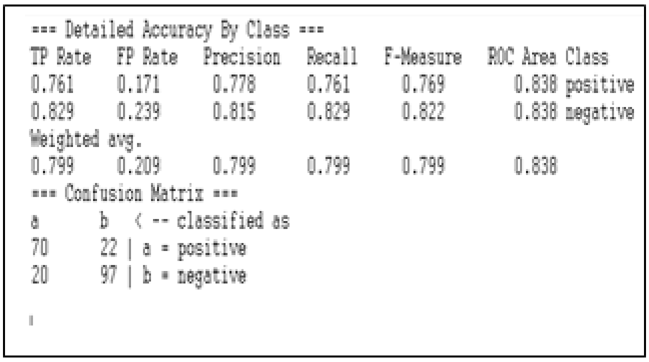


```
TP rate = TP/ (TP+FN)   ex :- (70) / (70+22) = 0.762 => Positive
                            (97) / (20+97) = 0.829=> Negative
FP rate = FP/ (FP+TN)    ex :-  (20) / (20+97) = 0.171 =>Positive
                              (22) / (22+70) = 0.239 =>Negative
Precision = TP/ (TP+FP) ex :-  (70) / (70+22)=0.778 =>  Positive
                              (22) / (22+97)=0.815 => Negative
Recall = TP/ (TP+FN) ex :-     (70) / (70+22) = 0.762 => Positive
                              (97) / (20+97) = 0.829=> Negative
F-Measure = (2*recall*precision)/(recall + precision)
Ex :- (2* 0.762*0.778) / (0.762 + 0.778) = 0.769 => Positive
         (2* 0.829*0.815) / (0.829 + 0.815) = 0.822 => Negative
*Note
|       TP=70    * Recall= TP rate
|       FN=22
|       FP=20
|       TN=97
```

i.e. appeared in figure 3. It includes the data of dataset analysis, for example, data about aggregate occurrences, arranged and unclassified cases, characterization exactness measures, confusion matrix and detailed precision.[20]

The above figure shows the relation information including name, instances, and attributes.[21] The 10 fold cross validation is representing the fact that the dataset is split into 10 different fold equally and then the algorithm is applied to each fold separately with different parameters.

This figure 2 shows the decision tree generated by the classifier, it includes the root nodes, the leaf nodes and the child nodes with their possible predicted values. For example, there are the number of leaf nodes are 5 and the size of a tree is 3

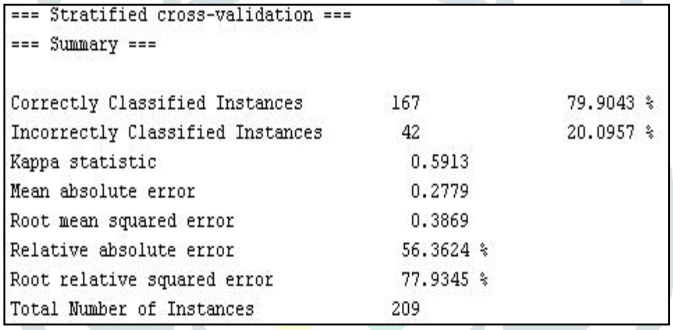


```
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area Class
0.761     0.171     0.778       0.761    0.769        0.838 positive
0.829     0.239     0.815       0.829    0.822        0.838 negative
Weighted avg.
0.799     0.209     0.799       0.799    0.799        0.838
=== Confusion Matrix ===
a       b   <-- classified as
70      22 | a = positive
20      97 | b = negative
```

Figure3: Final Result for detailed Accuracy

This above figure demonstrates the detailed information and confusion matrix delivered based on classes in the dataset. The classification technique will be performed with the use of PRM algorithm for mining technique. This particular rule has been extended and the resulted approach is known as CPAR technique. It works in manner when foil algorithm has been applied it generally deletes those records which are satisfactory to the rules parameter. In meanwhile for further processing, CPAR applied an extended version to give more accuracy to the analysis. The numbers of rules generated using hybrid approach should be greater than the FOIL generated results then the classification is supposed to be better.

Stratified Cross-Validation: Stratified Cross-Validation has been consisting of many important classification accuracy measures like kappa statistic, Mean Absolute Error, RMSE i.e. Root Mean Squared Error. The accuracy factors of algorithms have been shown only effectively by using these accuracy measures. [21]



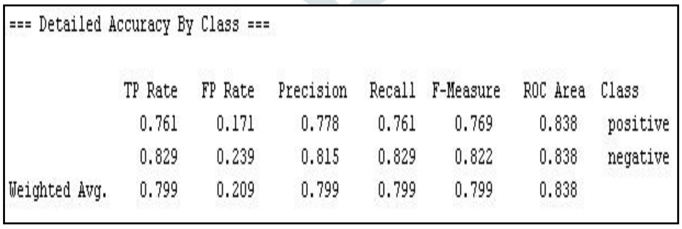
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         167               79.9043 %
Incorrectly Classified Instances        42               20.0957 %
Kappa statistic                          0.5913
Mean absolute error                      0.2779
Root mean squared error                  0.3869
Relative absolute error                 56.3624 %
Root relative squared error             77.9345 %
Total Number of Instances              209
```

Figure 4: Stratified Cross-Validation

In above figure, there are some essential measures that measure the accurately and erroneously characterized example in a dataset called mistakes. Mistakes are figured by the predefined equations which are talked about already. [22]

Detailed Accuracy by Class: The detailed exactness measures like the TP rate, the FP rate, Precision, Recall, F-Measure and ROC region are evaluated by class i.e. whether it is positive and negative which are coronary illness dataset's properties.



```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                 0.761    0.171     0.778    0.761    0.769      0.838   positive
                 0.829    0.239     0.815    0.829    0.822      0.838   negative
Weighted Avg.    0.799    0.209     0.799    0.799    0.799      0.838
```

Figure 5: Detailed Accuracy by class

The above figure shows the detailed result by class. The accuracy of the result is very much important as far as better classification is concerned.

VI. CONCLUSION AND FUTURE WORK

The paper discusses the numerical rigors and proficient and significant examination of bigger datasets to get at some particular outcomes. We have dissected the information of diseases such that the certainty support and confusion matrix made with learning & meaningful retrieval. The FOIL calculation improves grouping and results similarly as greater datasets are concerned. Fundamentally true informational collections have been taken with associated properties. The future work can be viewed as the FOIL calculation with the combination of various approach keeping in mind the end goal to enhance the accuracy of the class comes about. The future work can conspire of much efficient rules in which all the available records can be classified in one of the available classes. In meanwhile post-processing step can be added dynamically in the algorithm which will probably increase the classification accuracy.

REFERENCES

Augmented Naïve Bayes", ADMA, pp 186-194, 2015. DOI 10.14445/22312803/IJCTT-V32P105

[2] Siri Krishnan Wasan, Vasutha Bhatnagar, and Harleen Kaur "The Impact of Data Mining techniques on medical diagnostics", Data Science Journal, Vol 5(19), pp 119-126, October 2016.DOI 10.14445/22312803/IJCTT-V32P105.

[3] Shailza Chaudhary, Pradeep Kumar, Abhilasha Sharma, Ravideep Singh, "Lexicographic Logical Multi-Hashing For Frequent Itemset Mining", International Conference on Computing, Communication and Automation (ICCCA2015)

[4] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan, Yong Ren,"Information Security in Big Data: Privacy and Data Mining", 2014 VOLUME 2, IEEE 29th International Conference on Information Security in Big Data

[5] O.Jamsheela, Raju.G, "Frequent Itemset Mining Algorithms: A Literature Survey", 2015 IEEE International Advance Computing Conference (IACC)

[6] Feng Gui, Yunlong Ma, Feng Zhang, Min Liu, Fei Li, Weiming Shen, Hua Bai, "A Distributed Frequent Itemset Mining Algorithm Based on Spark", Proceedings of the 2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)

[7] Hongjian Qiu, Yihua Huang, Rong Gu, Chunfeng Yuan, "YAFIM: A Parallel Frequent Itemset Mining Algorithm with Spark", 2014 IEEE 28th International Parallel & Distributed Processing Symposium Workshops

[8] Paladugu S (2010) Temporal mining framework for risk reduction and early detection of chronic diseases. The University of Missouri-Columbia.

[9] Obenshain MK (2004) Application of data mining techniques to healthcare data. Infection Control and Hospital Epidemiology 25: 690-695.

[10] Shillabeer A, Roddick JF (2006) Towards role based hypothesis evaluation for health data mining. Electronic. Journal of Health Informatics 1: 1-9.

[11] Porter T, Green B (2009) Identifying Diabetic Patients: A Data Mining Approach.

[12] Panzarasa S, Quaglini S, Sacchi L, Cavallini A, Micieli G, et al. (2010) Data mining techniques for analyzing stroke care processes. In the Proc. of the 13th World Congress on Medical Informatics.

[13] Li L, Tang H, Wu Z, Gong J, Gruidl M, et al. (2004) Data mining techniques for cancer detection using serum proteomic profiling. Artificial intelligence in medicine 32: 71-83.

[14] Das R, Turkoglu I, Sengur A (2009) Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications 36: 7675-7680.

[15] Lakshmi K, Krishna MV, Kumar SP (2013) Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability. International Journal of Scientific and Research Publications 3: 1-10.

[16] Centers for Disease Control and Prevention (2013) Chronic Disease Prevention and Health Promotion. Accessed 27 September 2013, from http://www.cdc.gov/nccdphp/

[17] U.S department of health and human services (2005) High Blood Cholesterol What you need to know.

[18] Department of Health & Aging AG (2012) Seniors and Aged Care Australia websites have been replaced.

[19] Heller RF, Chinn S, TunstallPedoe HD, Rose G (1984) How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. British Medical Journal 288: 1409-1411.

[20] Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, et al. (1998) Prediction of coronary heart disease using risk factor categories. Circulation 97: 1837-1847.

[21] Simons LA, Simons J, Friedlander Y, McCallum J, Palaniappan L (2003) Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia 178: 113-116.

[22] Shahwan-Akl L (2010) Cardiovascular disease risk factors among adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing 1: 1-7.

[23] Dr.Neeraj Bhargava, Aakanksha Jain, Abhishek Kumar, Dac-Nhuong Le, (2017) pages 35 – Detection of Malicious Executables Using Rule-Based Classification Algorithms DOI: http://dx.doi.org/10.15439/978-83-949419-2-5