International Journal of Computer (IJC)

ISSN 2307-4523 (Print & Online)

© Global Society of Scientific Research and Researchers

http://ijcjournal.org/

Predictive Human Resource Analytics Using Data Mining Classification Techniques

Zarmina Jaffar[a*], Dr. Waheed Noor[b], Zartash Kanwal[c]

[a,b,c]*Computer Science and Information Technology, University of Balochistan Quetta, Pakistan*

[a]*Email: zaufi adii 89@yahoo.com*, [b]*Email: a.waheed.noor@gmail.com*, [c]*Email: zartash.cs@uob.edu.pk*

Abstract

The turnover ratio of employees in organization is the most important concerns as employees switching of organization/ job leaves huge gape and affects the performance of that organization. Among many, job satisfaction is the prime reason of employees to quit/switch, which is also directly related to human resource management (HRM) practices of the organization. It is always difficult and sometime beyond the control of human resource (HR) department to retained their well-trained and skilled employees but Data mining can play role to predict those employees who are expected to quit/leave an organization such that the HR department can device intervention strategy or look for alternative. In this paper, we focus on similar problem, where we use data mining techniques such as J48, Naive Bayes, and Logistic Regression predict employees who will leave the organization. Our data consists of different indicator values and some other important features such as number of projects, supervisor evaluation score and experience. We show that J48 perform well with accuracy 98.84% and TP rate 0.984%. Conventional statistical analysis has been used in literature to identify important factors affecting employees satisfaction but there is not agreed set. We also apply data mining techniques to identify such factors using two approaches such as Bayesian Network and IR. Finally, we provide a decision tree based model for decision makers that can easily stimulate employees satisfaction level for better retention policy.

Keywords: Human Resource Analytics;Human Resource Management,HR departments;Data Mining; Employee’s Performance.

1. Introduction

Data mining is the method to discover patterns from large amount of data by using different procedures. This is used as an analyzer for knowledge discovery databases used in decision making process.

----

* Corresponding author.

Large organizations employ it chiefly to discover new methods to raise their profits and to reduce the cost. Data mining analyze the data and helps to bring up the hidden factors so that useful patterns and information can be generated. These kinds of findings is definitely help any organization to take future decisions in relation to that product. There are several classification techniques in data mining such as the decision tree, neural network, rough set theory, baisian theory and fuzzy, see Phyu [1]. Decision takes place between trees, popular classification techniques, which make interpretable rules, or logic statement, See Jantan and his colleagues [2]. The rules can be used to predict the future through the predominantly established method. Data mining stands out due to its detailed procedures. Various domains such as statistics, artificial intelligence, mechanical algorithms, database systems, and visualization. Perform these effects on the basis of its applications for the business for which human resources management is clearly organized. Due to this, data mining has been found in popularity. Device with the ability to identify and pass trends within data Knowledge with predominantly predictive qualities, See Witten and his colleagues [3], Kurgan and his colleagues [4]. H. Liu and his colleagues [5] proposed a consistency based facility selection system. This happens when training instances are projected on a subset of attributes, then the value of the subset is evaluated from the level of stability in the values of the class. Consistency of any subset can never be less than the full set of attributes; Therefore, general practice is to use this subset evaluator in combination with random or excessive search, which seeks the smallest subgroup with the same consistency set attributes. M. Hall [6] proposes a new correlation based approach to address Correlation based feature selection (CFS) based on dependencies in a single dataset, and demonstrates how to apply the classification and regression of machines to both issues. Anirut Suebsing and his colleagues [7], presented Euclidean distance measurements to be used as a KDD dataset selection score. A high- score attribute greater than the defined threshold, selected as the best subset of features. Zahra Karimi and his colleagues [8], introduces a hybrid feature selection method combining symmetric uncertainty measurement and benefit measurement.According to the average fractional value calculated earlier, the SU and gain characteristics of each type of correlation feature are sorted. A high-level feature was selected within a range. He used the KDD dataset and the nave basic algorithm to evaluate his system. A. Chaudhary and his colleagues [9] provides a bayes of simple performance and a custom mobile device evaluation on three feature selection methods. In this task, the relevant methods, the gain ratio method and the information gain method are used. Ahmad and his colleagues [10] discussed that HRM (Human Resource Management is the major coalition of administration which deals with the most previous resource of the firm which is human resource. HRM educate their employee which can perform different task for the betterment of the people, society, surrounding and businesses. GHRM (Green Human Resource Management) is also the part of HRM. In an organization the GRM rules implemented through which their employees were upgraded to improve their profession which can help them to accomplish the firms ambition in a better way. Savaneviciene and his colleagues [11] described that Human Resource (HR) is the most valuable resource of any firm. To enhance the performance of organization the employees of an organization can follow the practices and methodologies of HRM. Most of the organization also select the software system which can efficiently control the day to day conditioning and necessities of HR department. Barney and his colleagues [12] conducted that better skill, encouragement, refine Human Resource Management carry out organizations constancy and as well as job gratification in which the firm should estimate and select to manipulate HR that make their employees stronger faith in the firms ambitions and valuation. Jantan and his colleagues [13] described that Human Resource Management is responsible for the selection of employees in organization. Human Resource professional have to decide the correct employee for the task at the correct time at the correct location. Arulrajah and his colleagues [14] discussed that by using Data Mining (DM) classifiers and classification method is very important to predict the future behavior of employees through past record sets. Arulrajah and his colleagues [15] conducted that Human Resource Management can help the organization to demonstrate a wellfounded administration in the association by upgrading their employees, agreement, to accept responsibility, transparency, effectual and efficacy, impartial and inclusive and also follow the rules and regulation. Jiang and his colleagues [16] conducted that Human Resource Management ordinary serve the honorable persons and also upgrade their honorable persons because insufficient recompense salary changeableness can also the main reason of the organizations ethical climate. Batt and his colleagues [17] In HRM 3 proportion for example (skillfulness, encouragement, and upgradation) were appreciative connected to humans central and retainer encouragement. Encouragement in Human Resource runover, for example, recompense them according to interpretation, motivation, benefaction, task protection are another probable to give retainer with external encouragement. Glebbeek and his colleagues [18] discussed that the observation displayed that turnover can have an unfavorable consequence on organizations interpretation. The Human Resource Management should have a conception the optimum estimation of the turnover is for their firms. In this paper, we build data mining model for Human Resource Analytics, where we use the employees record as our training dataset. The first stage, is to train different data mining models from the training data set using a suitable validation method. The learned model can be used for prediction on future/unseen data in the second stage**.**

2. Proposed methodology

A. Data Set

The primary data and the basic information on Predictive Human Resource Analytics collected from Kaggle website, by Vivek Aggrawal (Software engineer at Tata Consultancy Servies Gurugram, Haryana, India), for the purpose of evaluation by the selected method. In addition, data is converted into ARFF (Attribute Relation File Format) for processing in WEKA. The data used in this proposed study includes 14,999 observations, with each row representing one single employee. Fields in the dataset include the following 10 variables, See Figure 1.

The data set prepared ,pre-process and clean using the preprocess tab of the explorer window of the WEKA. The preprocessing capabilities of WEKA is recapitulate in expansive set of routines which is known as filters. The preprocess of a filters basically focus on two kinds of values such as instance and attribute in which the data set used to resample the values. The instances is divided into two datasets: a training set and a test set. 70% of the dataset selected for the training and 30% for testing and correcting the invert section have been selected to check the ratio of two values.

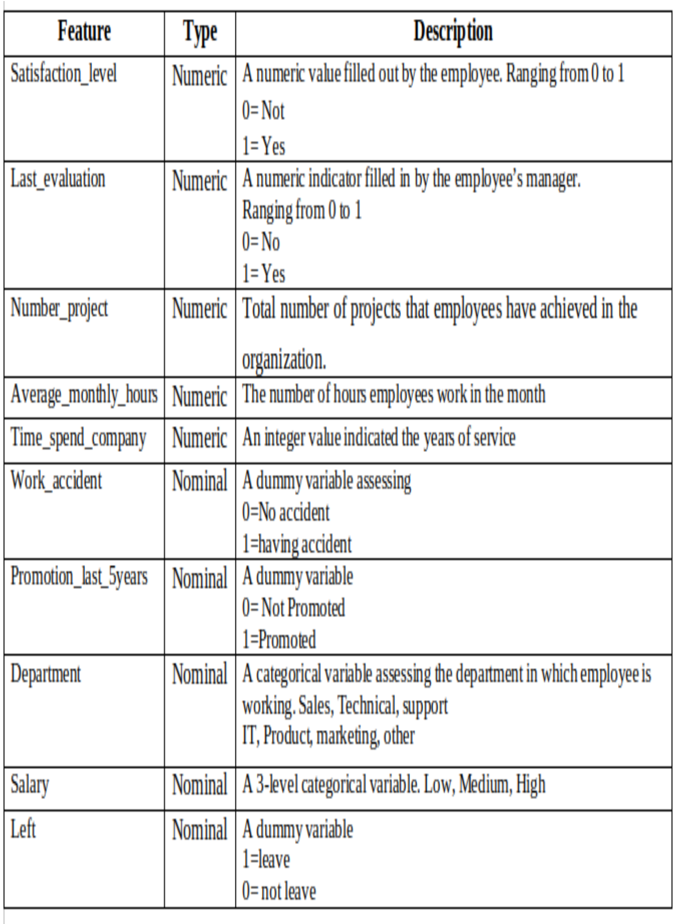
B. Data Pre-processing

The data which is available for mining is raw data, it is the original data Data can be in different formats, it comes from different sources, noise data can be irrelevant, data must be preprocessed before applying data mining techniques. The data mining algorithm uses the following steps:

1) Integration of Data: If the data comes from many different sources, then the data must have different aggregations, including the removal of the incompatibilities between datasets of different properties or attribute property values between sources of datasets.

2) Discretization: When the data mining algorithm can not face continuous characteristics, then discretization is required to be implemented. In this phase, changing a continuous attribute into categorical attribute, only to take some discrete values. Discretization often improves the understanding of searched knowledge.

3) Attribute Selection: All features are related, so selecting a subset of the features associated with them requires mining all the properties in the selected feature.



Feature	Type	Description
Satisfaction_level	Numeric	A numeric value filled out by the employee. Ranging from 0 to 1 0= Not 1= Yes
Last_evaluation	Numeric	A numeric indicator filled in by the employee's manager. Ranging from 0 to 1 0= No 1= Yes
Number_project	Numeric	Total number of projects that employees have achieved in the organization.
Average_monthly_hours	Numeric	The number of hours employees work in the month
Time_spend_company	Numeric	An integer value indicated the years of service
Work_accident	Nominal	A dummy variable assessing 0=No accident 1=having accident
Promotion_last_5years	Nominal	A dummy variable 0= Not Promoted 1=Promoted
Department	Nominal	A categorical variable assessing the department in which employee is working. Sales, Technical, support IT, Product, marketing, other
Salary	Nominal	A 3-level categorical variable. Low, Medium, High
Left	Nominal	A dummy variable 1=leave 0= not leave

Figure 1: Flow of proposed model.

C. Feature Selection

Many unrelated features appear in the data that will be mined. Therefore, it needs to be deleted. Many mining algorithms do not perform well when they have a large number of attributes or features. Therefore, before you apply, select the implementing to select any type of mining algorithm. The major reason of feature selection is to avoid over-fitting, refine model efficiency, and provide faster, more cost-effective products. The technique we used for feature selection are explained in following section, See Figure 2.

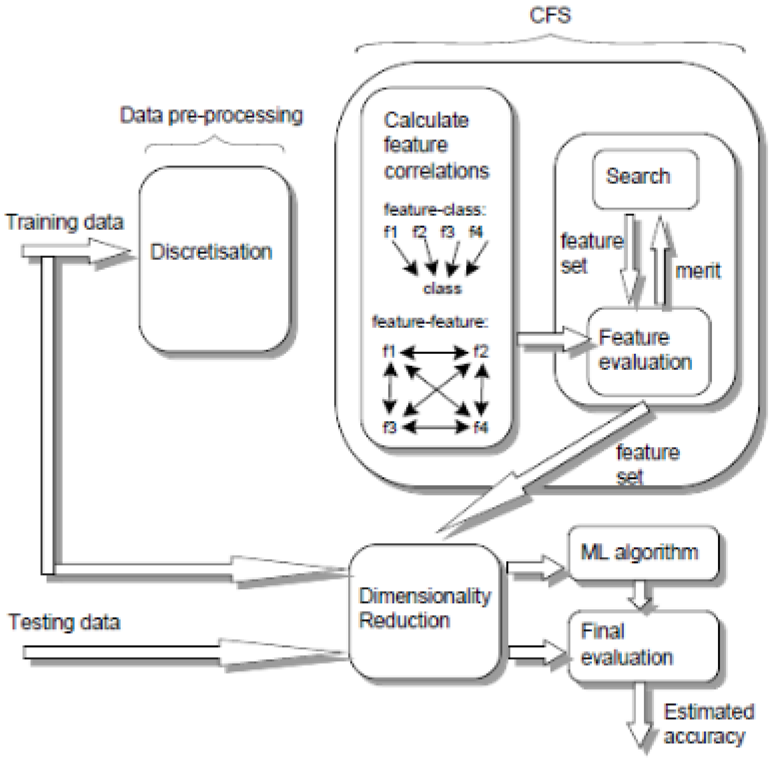


Figure 2: The components of CFS. Training and Testing data is reduce to

contain only the features selected by CFS Choosing the best feature adds an additional complexity to the modeling layer, rather than finding the best parameters for the complete set, finding the first best feature subset and including the model to customize. Attribute selection method can be roughly divided into filter and wrappar methods. The attribute selection method in the filter view is free to select from the data mining algorithm and evaluate the dependencies of the features and view only the internal properties of the data. In most cases, feature dependencies are calculated and low scoring features are removed. The observation method becomes very slow compared to the filter method because the data mining algorithm assumes that each property is applied to a subset. In addition, if many different data mining algorithms apply data, the calculation cost of the case method will be higher. Advantages of wrapper methods include interaction between model search and feature subsets in selection, and remembering feature dependencies. Another feature selection technique, called ”embedded technology”, is proposed, in which the best subset of features in the classifier construction process is found and can be considered searchable in a subset of facilities and a hypothetical combination location. As with the wrapper method, the embedded method is specific in a given learning algorithm. The advantage of embedded systems is that they involve interaction with the classification model, while the wrapper method is used at the same time. The feature selection techniques include CorrelationAttributeEval which requires a combination of the Ranker Search method. The primary focus of search behind the use of correlation-based method is to ensure that the connection between the variable and components in the testing can be estimated, which can be predicted by using the following formula.

$\mathrm{rzc} = \frac{\mathrm{krzi}}{\mathrm{k + k(k \square 1)rii}}$;

The rzc determines the correlation between the outside variable and the summed components where the number of components is represented by k and average of the correlation factor between the outside variable and the summed components. Lastly, determines the component intercorrelation. The formulae entail that the enhancement of the correlation factor concerning the external variable and the components increases the correlation between the outside variable and the composite. This factor is true even in the other situation where the correlation factor has been lower. This particular feature method used as the supervised classification task subset as a heuristic measure. It enables the reduction of the redundant attributes. The involvement of a supervised task is done by the participation of any feature other than numeric. That is why the attributes were further converted into the nominal using the discretization procedure in the preprocessing section. This particular aspect was utilized to enhance the prediction of the results by reduction in the redundant attributes and features. However, certain techniques and learning algorithms are also required that the performance of feature selection has been enhanced, which will be discussed in the wrapper section.

1) Correlation based Feature Selection: Feature selection is a preprocessing step to machine learning which is effective in reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility, See Yu **a**nd his colleagues [19] Steps of Feature Selection: If the characteristics of a subset are highly correlated with the class, but there is not much correlation with other properties of the class, the characteristics of the subset are good, See Hall and his colleagues [20]

Steps:

1) Subset generation: We have used four classifiers to rank all the features of the data set.Then we have used top 3, 4, and 5 features for classification.

2) Subset evaluation: Each classifier is applied to generated subset.

3) Stopping criterion: Testing process continues until 5 features of the subset are selected.

4) Result validation: We have used 10-fold cross validation method for testing each classifiers accuracy.

2) Information Gain Feature Selection: Gain ratio is another evaluator indicator used in feature selection evaluation, See formulas in equation (1). A higher gain ratio value shows a greater correlation between properties relative to class labels. GainR(Class,Attr) = (H(Class)H(Class—Attr))/H(Attr) (1) where H specifies the entropy (information). Gain ratio AttributeEval method used in Weka as a feature selection method. This yield ratio calculates the value of the in gain ratio relative to the scale of the class. After 10 fold cross-validation after each property is evaluated separately, the ”search” method attributes the ranker method property to ranker through its separate evaluation.

D. Classification

Data mining algorithms have three different learning methods: supervised, unsupervised, or semi-supervised. In the supervised learning, the algorithm works with groups of examples whose labels are known. In the case of classification work, the label may be a nominal value, or in the case of regression, the numeric value can be. On the contrary, the examples of labels in the dataset are unknown, and algorithms usually have examples of their attribute values. Targeting the group according to equality clustering work. Classification tasks can be seen as supervised techniques, where each instance corresponds to a class, represented by the value of a particular target feature or only a class property. The knowledge detected by classification algorithm can be expressed in many ways, such as rules, decision tree, bayesian network and so on. The different techniques we use for classification are explained in the following section, See Figure 3.

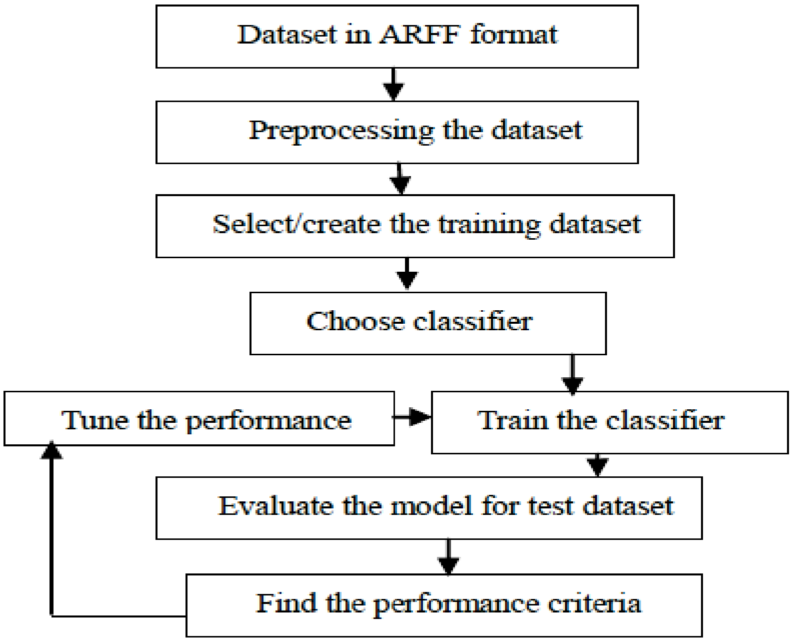

Dataset in ARFF format
Preprocessing the dataset
Select/create the training dataset
Choose classifier
Tune the performance
Train the classifier
Evaluate the model for test dataset
Find the performance criteria

Figure 3: Classification step through trainig and testing data.

The classification process is divided into two stages: training, when the training model is created from the training set, the model is evaluated in the test set. At the training stage, all examples of predictive characteristics and training sets in the algorithm have access to the values of the target attribute, and it uses that information to create a classification model.

The classification process is divided into two stages: training, when the training model is created from the training set, the model is evaluated in the test set. One of the key goal of the classification algorithm is to maximize the accuracy of the classification through the classification model, while predicting the examples in the tests planned during the receiving training.

E. Association

Association is considered to be one of the main technique in data mining applications. It reveals all the interesting relationships that might be called associations in a large database, See Figure 4.

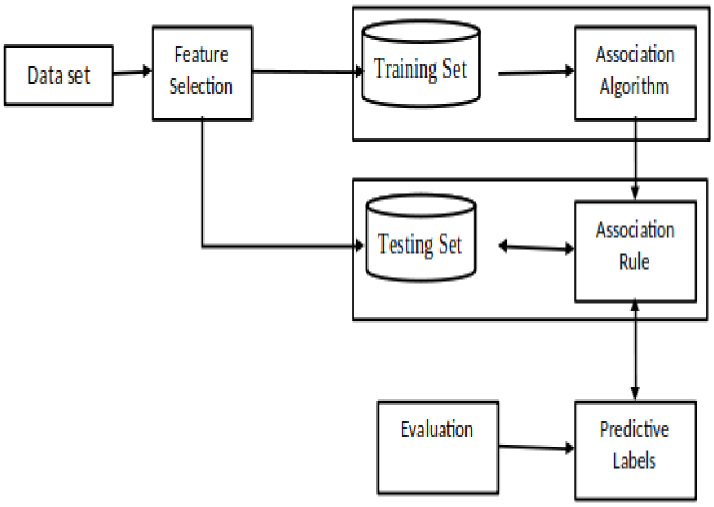


Figure 4: Proposed Association model.

1) Apriori: Based on the rules of association to find relations between different objects. We first need to find the item set in the data set and analyze these item sets.

We make association rules and then evaluate the decision. According to these rules, the data, finally, select the rules that have larger confidence and support than required smallest one (Piatetsky- Shapiro, 1991).This is usually used in the decision support area. In my data set, we can also use association rules. But, the decision is a better structure of the tree that we can get the rules more clearly.

3. Experiment i

The proposed experiments are capable of decision tree classification approaches to distinguish the aptitude design in the subset of Human Resource DB and different decision trees and classification approaches are used for the given datasets. In this article different classification techniques are used for talentd dataset that focus on the accuracy of datasets. Different classification approaches depend on the input vaiables.

To develop the prediction model, the classification techniques were used. As we have discussed about decision tree it was also used for prediction and machine learning and neural network etc.

There are different types of classification techniques such as J48,Nave Bayes, Bayes Net, Logistic, and OneR, Jrip, Random Tree, SVM etc. To check the accuracy of classification technique we apply 10 folds of cross validation on data set. As we have discussed about two types of data sets training sets and testing sets.

In the Explorer tab we just select the train data set to access the classify tab, than load the model which we have previously save and apply different classification techniques to check the accuracy of each classifiers.

In table I the accuracy for full attributes have been given in which J48 will be the best among all the classifiers.

TABLE I: classification results on different classifier using 10 fold cross validation.

Classifier Algorithm	Data Set %
J48	98.39
Naive Bayes	93.75
Bayes Net	84.69
OneR	89.45
Logistic Regressions	79.23

1) J48: J48 is one of the best and widely used technique with high accuracy generally used as a supervised learning methods. It is basically used for prediction and classification techniques and very easy to understand.

TABLE II: classification results on j48 classifier using 10 fold cross validation.

Classifier Algorithm	Training Set (%)	Testing Set (%)
J48	98.84	97.67
Navie Bayes	85.35	84.53
Bayes Net	93.83	93.18
OneR	89.35	89.45
Logistic Regressions	79.75	80.21

2) Naive Bayes: Naive Bayes basically based on the theorem of Baye. This is the most popular classification technique due to its simplicity, computing efficiency and excellent performance for real-world challenges. This train, evaluates data very quickly with much greater accuracy.

3) Bayes Network: This is a visual model that can define the set of variable conditional independences.

4) Logistic: Logistic regression is the most important modelling tool. It is used to predict the analysis and is used in projects because the response variables logistic regression is considered as a powerful modeling tool. Logistic regression is the correct regression assessment, when structured variables are dicotomas (binary). Like any regression analysis, logistic regression is a predictive analysis. Logistic regression was considered in the character of the reactive variable discrete character. During the prediction, it creates a version to expect the likelihood of its occurrence.

5) OneR: Comparing classification techniques with other models is fundamental. OneR classifier originally produces a one-degree selection tree expressed as a set of guidelines, which everybody examines a special feature. It is very simple, reasonably priced technique that produces regularly among the top policies with high accuracy for assessing other classification fashion, and predictive energy of specific characteristics as a hallmark of the energy. In these experiments selected classifiers from decision tree algorithm are used to check the accuracy of different classification approaches. The accuracy for full attributes have been given in table I and accuracy for training set and testing set have been in table II. In all experiments J48 considered to be the suitable algorithm among all the others.

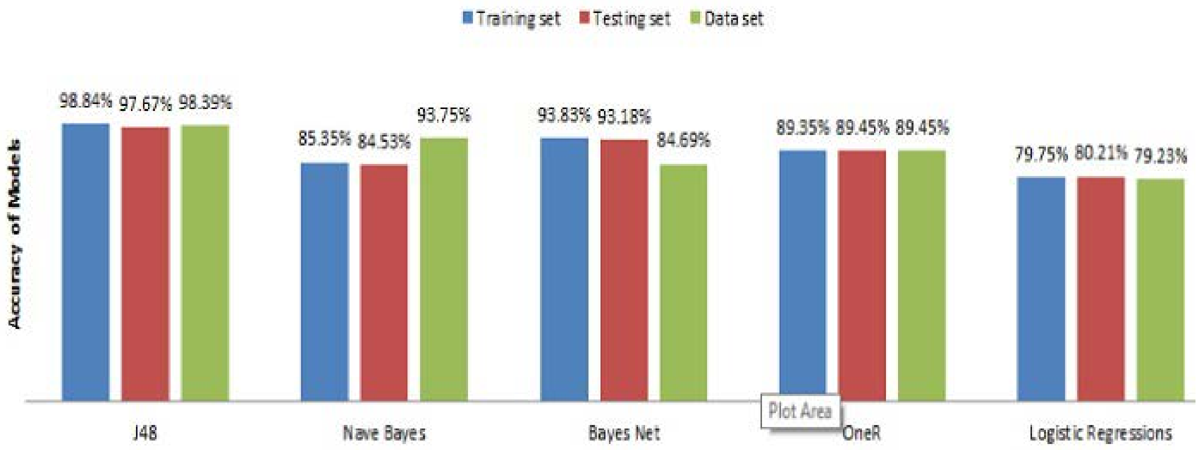


Figure 5

4. Conclusion

We compare different feature selection methods which are nproposed and tested on secondary data. Naive Bayes, Logistic Regression, and J48 classifier are shown in this paper along with the methods of selection of various features. Existing feature selection algorithms may not be able to generate a valid subset of features for the classification of many different regions. Although some algorithms may reduce features, their classification accuracy is not high. The proposed information selection algorithm based on conditional equivalence produces high efficiency and small characteristics in several different data sets, and the classification accuracy is higher. Our proposed algorithm will be improved by eliminating the relevant and effective functions of the machine

References

[1] T. N. Phyu, “Survey of classification techniques in data mining,” in Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, 2009, pp. 18–20.

[2] H. Jantan, A. R. Hamdan, and Z. A. Othman, “Human talent prediction in hrm using c4. 5 classification algorithm,” International Journal on Computer Science and Engineering, vol. 2, no. 8, pp. 2526–2534, 2010.

[3] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

[4] L. A. Kurgan and P. Musilek, "A survey of knowledge discovery and data mining process models," The Knowledge Engineering Review, vol. 21, no. 1, pp. 1–24, 2006.

[5] H. Liu, R. Setiono et al., "A probabilistic approach to feature selection-a filter solution," in ICML, vol. 96. Citeseer, 1996, pp. 319–327.

[6] M. A. Hall, "Feature selection for discrete and numeric class machine learning," 1999.

[7] A. Suebsing and N. Hiransakolwong, "Euclidean-based feature selection for network intrusion detection," in International Conference on Machine Learning and Computing, vol. 3, 2011, pp. 222–229.

[8] Z. Karimi, M. M. R. Kashani, and A. Harounabadi, "Feature ranking in intrusion detection dataset using combination of filtering methods," International Journal of Computer Applications, vol. 78, no. 4, 2013.

[9] N. A. A. Shashoa, N. A. Salem, I. N. Jleta, and O. Abusaeeda, "Classification depend on linear discriminant analysis using desired outputs," in Sciences and Techniques of Automatic Control and Computer Engineering (STA), 2016 17th International Conference on. IEEE, 2016, pp. 328–332.

[10] S. Ahmad, "Green human resource management: Policies and practices," Cogent Business & Management, vol. 2, no. 1, p. 1030817, 2015.

[11] A. Savaneviciene and Z. Stankeviciute, "Human resource management practices linkage with organizational commitment and job satisfaction." Economics & Management, vol. 16, 2011.

[12] J. B. Barney and P. M. Wright, "On becoming a strategic partner: The role of human resources in gaining competitive advantage," Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management, vol. 37, no. 1, pp. 31–46, 1998.

[13] H. Jantan, A. Hamdan, Z. Othman, and M. Puteh, "Applying data mining classification techniques for employee's performance prediction," in Knowledge Management 5th International Conference (KMICe2010), 2010, pp. 645–652.

[14] A. A. Arulrajah, "Literature review on good governance in the organizations through human resource management: A corporate level analysis," International Business Research, vol. 9, no. 8, p. 14, 2016.

[15] "Contribution of human resource management in creating and sustaining ethical climate in the organisations," Sri Lankan Journal of Human Resource Management, vol. 5, no. 1, 2015.

[16] K. Jiang, D. P. Lepak, J. Hu, and J. C. Baer, “How does human resource management influence organizational outcomes? a meta-analytic investigation of mediating mechanisms,” Academy of management Journal, vol. 55, no. 6, pp. 1264–1294, 2012.

[17] R. Batt and A. J. Colvin, “An employment systems approach to turnover: Human resources practices, quits, dismissals, and performance,” Academy of management Journal, vol. 54, no. 4, pp. 695–717, 2011.

[18] A. C. Glebbeek and E. H. Bax, “Is high employee turnover really harmful? an empirical test using company records,” Academy of Management Journal, vol. 47, no. 2, pp. 277–286, 2004.

[19] L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” in Proceedings of the 20th international conference on machine learning (ICML-03), 2003, pp. 856–863.

[20] M. A. Hall, “Correlation-based feature selection for machine learning,” 1999.