New Approach from K-Nearest Neighbor Data Classification

[1]Ms.P.Anushya , [2]Ms.P.Jothilakshimi , [3]Ms.S.Pavithra , [4]Mrs.B.Revathi
[1,2,3]CSE dep , Mangayarkarasi College of engineering , Madurai, Tamilnadu.
[4]AP/ CSE dept , Mangayarkarasi College of engineering, Madurai, Tamilnadu.

Abstract—

Data mining, the extraction of masked predictive information from large database the comparative study between Decision tree Algorithm and K- Nearest Neighbor Algorithm of Classification techniques is present in this paper. .strikingly to ensure the appropriate person admeasure to the accommodating job at the right time. From here, the interest of data mining (DM) role has been flourishing that its objective is the discovery of wisdom from huge amounts of data. In this paper, DM techniques were utilized to build a classification model for predicting employees' performance using a real dataset collected real time employee from the through a canvass prepared and distributed for 145 employees. Three main DM techniques were used for building the classification model and identifying the most yielding factors that positively affect the performance. The proficiency are the Decision Tree (DT), Naïve Bayes, and Support Vector Machine (SVM). To get a highly accurate model, several experiments were executed based on the previous techniques that are implemented in WEKA tool for enabling decision makers and human resources professionals to predict and enhance the performance of their employees using the UCI magazine machine learning .It is used to elicitation useful knowledge regarding their need. Data mining has many techniques.

Index Terms — *Classification, Mining, Employees' Performance, HRM, Naïve Bayes, SVM, Decision Tree, Distance Metric.*

I. INTRODUCTION

HRM has a leading role in deciding the competitiveness and effectiveness for better continuation. Organizations consider HRM as "people practices". Therefore, it becomes the responsibility of the HRM to allocate the best employees to the appropriate job at the right time, train and qualify them, and build evaluation systems to monitor their performance and an attempt to preserve the potential talents of employees [1].

With the advancement and enlargement of technologies in business organizations, HR employees need not handle the massive amount of data manually any further. These data is very important for the decision makers, but there is a challenge to mine and get the best and useful data from these huge data [1]. From here, the role of DM comes. DM is a step in Knowledge Discovery in Database (KDD) and is currently acquiring great deal of attention and utilization. It is considered as a recently emerging analysis and predictive tool [2], because of the existence and multiplicity of massive amount of data containing huge hidden unknown knowledge. Knowledge can be extracted through various methods and one of them is by using DM technique.

DM techniques provides an approach to utilize different DM tasks such as classification, association, and clustering used to extract hidden knowledge from huge expance of data.

Classification is a predictive DM technique, makes prediction about values of data using known results found from various data. Classification technique is a supervised learning technique in DM and machine learning, whereas the class level or the target class is already previously known. It is one of the most useful tasks in DM to build classification models from an input dataset. The used classification techniques commonly build models, which in turn used to predict future data trends [3]. With classification, Predictive models have the specific bull's eye of enabling us to predict the unknown values of variables depending on interest previously known values of other variables [4].

In this connection, the main objectives of the present study were extracted to support the decision makers in different locations to discover potential talents of employees as follows:

- Gathering a dataset of predictive variables,
- Identification of different factors, which affects employees' behavior and performance
- Using proposed DM classification techniques for constructing a predictive model and identifying relationships between most important factors affecting over whole efficiency of the model

.

There are various data classification techniques such as DT, SVM, Naïve Bayes classifier, and others. In this paper, the classification process is executed through using the three main classification technique that were mentioned above. Other techniques can also be used for classification such as Neural Network (NN), K-Nearest Neighbors (KNN), etc.

II. DATA MINING AND DATA MINING TECHNIQUES

A. *Definition of Data Mining*

Data mining is a method that automates the detection of relevant patterns and relationships given a dataset. It uses defined approaches and algorithms to scan the given data set and predict the data trend. Data Mining is not particularly new. Statisticians have been using similar manual approaches to review data and propose a trend. However, what has changed is the volume of data today has increased immensely and thus giving rise to automated data mining techniques that investigate data trends very quickly. Users can also determine the outcome of the data analysis by the parameters they choose, thus automated data mining gives users such flexibility.

B. *Data Mining Techniques*

Data Mining has a broad spectrum of application and so there are may techniques of data mining in existence. Most commonly used techniques are Decision Trees and The Nearest-Neighbor method. Each of these techniques analyzes data differently and is discussed below.

- **Decision Trees:**

These are tree like graphs or models of decisions and their possible outcomes. They consist of flow-chart like structure that has nodes to represent a test on attributes and the branches of the node represent the outcome of the test. They are often used in data analysis as they depict every possible scenario pictorially and hence patterns can easily be identified in a data set.

- **Nearest Neighbor**:

This method classifies data based on a majority vote of its neighbors and it is assigned to the class most common amongst all its neighbors. This method is discussed in detail in later sections of this paper.

The k-Nearest Neighbor Algorithm involves two phases.

- The Training Phase
- The Testing Phase

These Phases are discussed in detail in the following subsections.

1) *The Training Phase*

kNN Algorithm does not explicitly require any training phase for the data to be classified. The training phase usually involves storing the data vector co-ordinates along with the class label. The class label in general is used as an identifier for the data vector. This is used to classify data vectors during the testing phase

2) *The Testing Phase*

Given data points for testing, our aim is to find the class label for the new point. The algorithm is discussed for k=1 or 1 Nearest Neighbor rule and then extended for k=k or

K Nearest nighbor rule.

a) k=1 or 1 Nearest Neighbor This is the simplest scenario for classification.

Let ‘x’ be the point to be labeled.

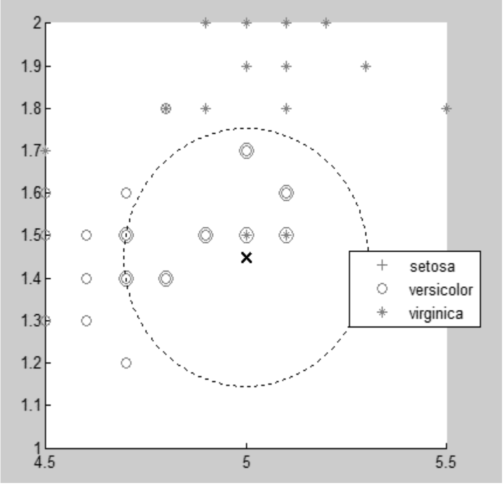
• Find the point closest to ‘x’ in the training data set. Let this closest point be ‘y’.

• Now nearest neighbor rule asks to assign the label of ‘y’ to ‘x’. This seems too simplistic and some times even counter intuitive. This reasoning holds only when the number of data points is not very large. If the number of data points is very large, then there is a very high chance that label of x and y are same

.

Distance Metrics

Distance metrics are a method to find distance between a new data point and existing training dataset. In this research, we experiment with 11 distance metrics, which can be explained as follows



dist_Minkowsky(A,B)=(∑i=1m|xi−yi|r)1/r

dist_correlation(A,B)=∑mi=1(xi−μi)(yi−μi)∑mi=1(xi

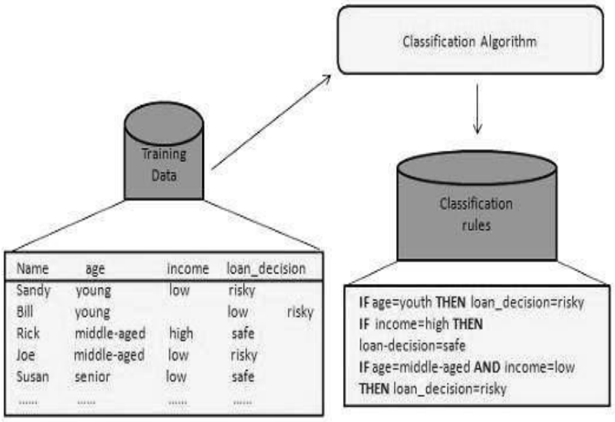
−µi)2∑mi=1(yi−µi)2——————————————————————√

dist_Chi-square(A,B)=∑i=1m1sumi(xisizeQ−yisizeI)2

Classification in Data Mining

Databases or data warehouses are rich with hidden information that can be used to provide intelligent decision making. Intelligent decision refers to the ability to make automated decision that is quite similar to human decision. Classification and prediction are some of the methods that can produce intelligent decision. Currently, many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics.

In this study, we are focusing on classification methods in data mining as part of machine learning process. Classification and prediction in data mining are two forms of data analysis that can be used to extract models to describe important data classes or to predict future data trends(Han & Kamber, 2006). The classification process has two phases; the first phase is learning process where the training data are analyzed by the classification algorithm. Learned model or classifier is represented in the form of classification rules. The second phase is classification process, where the test data are used to estimate the accuracy of classification model or classifier. If the accuracy is considered acceptable, the model can be applied to the new data to know the prediction result



Name	age	income	loan_decision	
Sandy	young	low	risky	
Bill	young		low	risky
Rick	middle-aged	high	safe	
Joe	middle-aged	low	risky	
Susan	senior	low	safe	
......	......	......	......	

IF age=youth THEN loan_decision=risky
IF income=high THEN
loan-decision=safe
IF age=middle-aged AND income=low
THEN loan_decision=risky

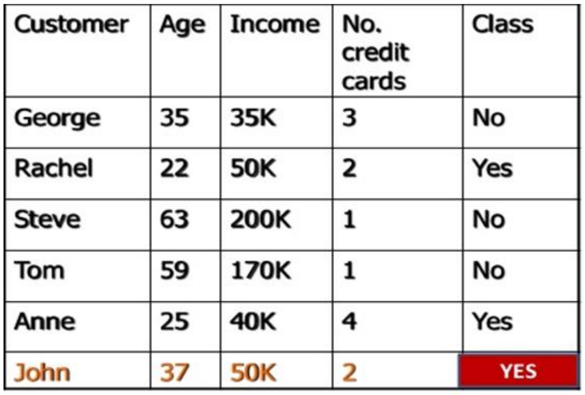
1. *Employee manager*

In this module the graphical user interface provides adding new employee details to the company. In this module, you will learn how to add, edit and delete the Employee's information from the database using java swing. For this purpose, we have used three tabs using JTabbedPane class. Under these tabs, we have created three forms using JLabel and JTextField class which are enclosed into a panel which is then added to the JTabbedPane.

2.Pay slip generator

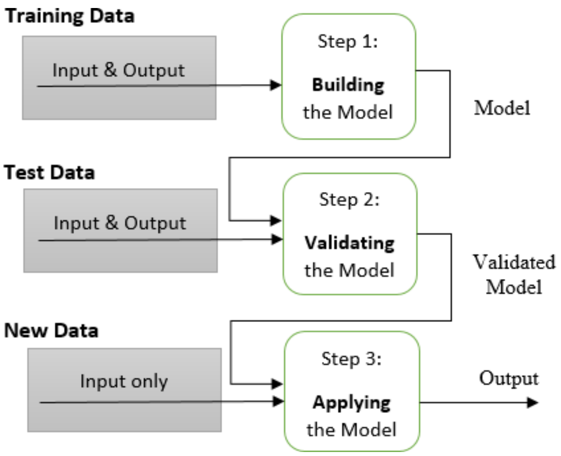
A Pay slip is a document issued to an employee that itemizes each component of earnings and deductions, and the net amount paid to an employee for a given pay period. It provides visibility to an employee of how the net amount has been arrived at. In this module it asks for employee id. When finance admin enters the employee id it displays the employee current details retrieved from the database these includes name, designation, department, salary and other personal details. Pay slip generator

A Pay slip is a document issued to an employee that itemizes each component of earnings and deductions, and the net amount paid to an employee for a given pay period. It provides visibility to an employee of how the net amount has been arrived at. In this module it asks for employee id. When finance admin enters the employee id it displays the employee current details retrieved from the database these includes name, designation, department, salary and other personal details.



Customer	Age	Income	No. credit cards	Class
George	35	35K	3	No
Rachel	22	50K	2	Yes
Steve	63	200K	1	No
Tom	59	170K	1	No
Anne	25	40K	4	Yes
John	37	50K	2	YES

In General, this paper is an initiative attempt to investigate DM tasks, especially classification task, for supporting decision makers and HR's professionals by identifying and studying the main factors of their employees that may positively affect their performance. The paper applied some of the classification techniques to build a proposed model for supporting the prediction of the employees' performance. In the next sections, a comprehensive description of the study is presented, specifying the methodology, the experiments and results, and a discussion of the results, finally conclusions and recommendations for future work.



CONSTRUCTING THE CLASSIFICATION MODEL

The proposed methodology was adopted for the objective, which is building the classification model studying certain factors that may affect and predict the employees' performance. For achieving this objective, it is necessary to exist a generic guide to develop a DM project lifecycle containing certain steps that includes Problem Definition and Objective Structuring, Data Collection and Understanding, Data Preparing and Preprocessing, Modeling and Experiments, Testing and Evaluating.

In general, Classification contains some steps to complete its process. The first step is called the learning step where in the model; predefined classes are built by analyzing a set of training dataset variables. Each variable is assumed that has a relation and regards to a predefined class. The second step is responsible for estimating the accuracy of model or classifier (validating the model) through testing the model using a different dataset. If the classifier's accuracy was considered acceptable, the model or classifier can be used to apply to new unseen data to give prediction about specific unknown label class and this is considered the third step as shown in figure 1. Therefore, the model acts as a classifier in the process of decision-making. There are various classification techniques have been used in the prediction process such as DT, Naïve Bayes, SVM, etc.

A. *Problem Definition and Objective tructuring*

The first step in data mining is to understand and define the right problem and specify the objectives. Meanwhile, data miners should also equip themselves with domain knowledge to understand problem nature, which will greatly improve DM effectiveness and efficiency. Indeed, human resource management activities are very complicated and thus few quantitative approaches have been employed in practice [2]. HRM at most of other public sectors use traditional assessment techniques that they do not enable them to get the perfect assessment for the employees' performance and therefore they cannot predict the performance and discover the talents.

this research concentrates on how can present a proposed model supporting HRM and Decision makers to predict the employees' performance of dataset and identifying the employees' factors that are affect and associate with bad/good performance. Moreover, detecting the most suitable DM technique with the most highly accuracy between the various classification techniques that will be used.

A. *Data Collection and Understanding Process*

The idea of this study is building a classification model for predicting the employees' performance based on a real dataset to get real and significant results for supporting the HR executives and the decision makers. To collect the required data, it is necessary to exist a practical way. Therefore, a questionnaire is prepared and manually distributed the employees of MOCA containing the several attributes that may affect and predict the performance Class (the target Class). The asked attributes for training dataset are selected based on the related factors for employee performance that confined between Educational factors, Personal factors, and Professional factors such as (job title, age, rank, qualifications, grade…etc.) as illustrated in table 1. These attributes are used to predict the employee performance (the target class) to be - Excellent, Very Good, or Good. The questionnaire was filled by 145 employees from all different sectors with various job titles, ages, and ranks to get complete sample about them.

B. *Data Preparation and Pre-processing*

After the process of questionnaire collection finished, the process of preparing the data is performed, the raw data contained instances that were not applicable. This was due to errors and anomalies that had to be discarded. The data was transferred to Excel sheets to review and modify the types of the collected data where some attributes types need to be changed from numeric data type into categorical data type i.e. values illustrated by ranges for example the attributes of No. of experience years and service period (X3, X4) according to table 1. Other attributes need to be generalized in fewer discrete values instead of that they already for example the attribute of faculty specialization (X15) according to table 1 contained values like IT, CS, MIS they have been considered as only one value, IS and so on. Therefore, Data generalization is also considered as one of the data reduction techniques. After preparing the excel sheet and making the needed processing, the file was transformed into arff format that is compatible with the WEKA DM toolkit which was used in building the model.

The WEKA (Waikato Environment for Knowledge Analysis) toolkit is a machine learning platform, developed by researchers at the University. Java is the used implementation language. It provides a unified package at only one application, which enables users to access the modern updated technologies in DM and machine-learning environment. It contains several tasks such as pre-processing, classification, clustering, association and visualization. The WEKA tool had an important advantage where it is available for free and has a simple GUI so, it could be used smoothly. The tools supported by the WEKA workbench are based on statistical evaluations of the models (algorithms). Consequently, the WEKA user can easily make comparisons among the results and accuracies of the applied machine learning and DM algorithms for a given dataset in flexible procedures in order to detect the most suitable algorithm for the given dataset .

Feature Selection

Feature selection is a one of the main concepts of (DM)Distance metric. Where, it is a process of selecting necessary useful variables in a dataset to improve the results of machine learning and make it more accurate. At which, Using too many numbers of variables in a dataset reduce predictive performance. The data set may contain too many features; some of them do not promote the prediction accuracy, and thus make the predictive model excessively complicated. Therefore, unnecessary useless variables must be avoided to make the model efficiently works. Deciding which unnecessary variable to avoid can be done by a manual manner using domain knowledge or it can be done automatically .

This paper targets getting the most important variables that may positively affect the accuracy of the employees' performance prediction model using the various feature selection algorithms that are supported in WEKA such as CorrelationAttributeEvaalgorithm, GainRatioAttributeEval algorithm, ReliefFAttributeEval algorithm, and so on.

No .	**Techniqu e**	**Prediction Accuracy**
1	*KNN*	77.93 %
2	*Naïve Bayes*	88.90 %
3	*SVM*	81.38 %

CONCLUSION AND FUTURE WORK

Applying the DM techniques in the different problem domains in the HRM field is considered as an important and urgent issue. Especially, at the public sector in Egypt. In addition, increasing the horizons of academic and practice research on DM in HR for reaching a government sector with a high performance.

This paper has concentrated on the capability of building a predictive model for employees' performance of using classification techniques through studying and testing the factors that might positively affect the performance of the employees. The SVM technique was found as the most suitable classifier for building the predictive model, where it had the greatest prediction accuracy through all the three experiments that had executed with the highest percentage 86.90%. WEKA toolkit was used through executing the experiments.

For decision makers and HRM department, this model, or an enhanced one, can be utilized in predicting the performance of the potential talents that will be promoted, predicting the performance of the recently applicant employees where various actions can be taken for avoiding any risk related to hiring employees with a low performance, or so on.

REFERENCES

[1] L. Sadath, (2013) "Data Mining: A Tool for Knowledge Management in Human Resource," International Journal of Innovative Technology and Exploring Engineering, Vol. 2, Issue 6, April 2013.

[2] G. K. Gupta (2006) "Introduction to Data Mining with Case Studies" ISBN-81-203-3053-6.

[3] AI-Radaideh, Q. A., AI-Shawakfa, E.M., and AI-Najjar, M. I., (2006) "Mining Student Data using Decision Trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[4] Surjeet K. Y., Brijesh B., Saurabh P., (2011) "Data Mining Applications: A comparative Study for Predicting Student's performance", International Journal of Innovative Technology and Creative Engineering, Vol.1 No.12 (2011) 13-19.

[5] Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010b). "Human talent prediction in HRM using c4.5 classification algorithm". International Journal on Computer Science and Engineering, 2 (08-2010), PP. 2526–2534 [D].

[6] Islam, M. J., Wu, Q. M. J., Ahmadi, M., and Sid-Ahmed, M. A., (2010), "Investigating the Performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers" Journal of Convergence Information Technology Volume 5, Number 2, April 2010.

[7] V.Kalaivani, Mr.M.Elamparithi (2014), "An Efficient Classification Algorithms for Employee Performance Prediction", International Journal of Research in Advent Technology, Vol.2, No.9, September 2014 E-ISSN: 2321-9637.

[8] S.Yasodha and P. S.Prakash, (2012), "Data Mining Classification Technique for Talent Management using SVM," the International Conference on Computing, Electronics and Electrical Technologies, 2012.

[9] Hua Hu, Jing Ye, and Chunlai Chai, (2009), "A Talent Classification Method Based on SVM", in International

Symposium on Intelligent Ubiquitous Computing and Education, Chengdu, China, 2009, pp. 160-163.

[10] Kirimi JM, Motur CA (2016), "Application of Data Mining Classification in Employee Performance Prediction". International Journal of Computer Applications, Volume 146 – No.7, July 2016.

[11] Desouki M. S., Al-Daher J (2015), "Using Data Mining Tools to Improve the Performance Appraisal Procedure, HIAST Case". International Journal of Advanced Information in Arts, Science & Management Vol.2, No.1, February 2015.