Chronic Kidney Disease Prediction using Supervised Classification Techniques

Ms. Veerpal Kaur [1,] Ms. Rupali Zakhmi[2]

Swami Vivekanand Institutes of Engineering & Technology, Banur (Chandigarh)

ABSTRACT--A set of algorithms is required in the balanced combination in order to achieve the goal related to the chronic kidney disease classification for online healthcare databases. The existing model has used to certain weight clustering in the second level to divide the data into the training and testing data before evaluating the performance of the classification algorithm. The new model has been proposed for the use of data scaling using the Min Max scaling along with missing value handling with mean value for the quantitative variables. For the case of qualitative or categorical data, the missing value handling is performed along with the numeric labeling of data followed by dummy variable creation. The k-nearest neighbor (kNN) and support vector machine (SVM) algorithms can be used to match and find the category of the target data entry. The KNN or SVM based classification algorithm will be used between the feature descriptors obtained from the test data rows and training data rows & categories. These descriptor vectors will be analyzed on the final stage and compare with the feature descriptor vectors obtained from the each data entry in the test and training matrices. Then the result would undergo the final matching and returns the category of the target row in the form of presence or absence of disease. The performance of the proposed algorithm will be measured on the basis of Precision, Recall, F1 measure, Accuracy and Time. The second observation is that all keywords devised so far have been designed to preserve as much as possible text content by working on small neighborhoods (or in combinative features) of the local variables. The SVM has been recorded with 98.92% (mean) and 99% (median) of accuracy, which is significantly higher than KNN's 97% (both mean and median). Also SVM outperformed KNN on the basis of precision by (98.37% mean) and recall (99.94% mean) against 96.95% (precision mean) and 98.18% (recall mean)**.**

KEYWORDS—Classification, Machine learning, KNN, SVM

I. INTRODUCTION

Information about the mining could be the technique of inspecting & gather info by different perspectives along with outlining the idea in beneficial details. The idea details helpful to improve income, reduces prices, or both equally. It really is used for gather the data by different- different web sites. Information about the mining could be the technique of finding info within big relational listings.The item a good choice for end users to handle the information effortlessly. Process for finding files habits undetectable inside significant files units. In other words Data mining features captivated a great deal of consideration inside the facts market and also inside modern society as a whole nowadays. Data mining is the term for removing or even "mining" understanding by a lot connected with files, typically immediately compiled.

- Data mining finds useful information hidden in large volumes of data. It is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. The following section shows the functionalities of the collected data to reach conclusive evidence:
- Collection of data, and designing the database for data ingestion
- Management of the data in the form of ingestion and processing pipelines
- Perform the advanced analytics on the ingested data, which is saved in the warehouse in the form of online analytical processing (OLAP) architecture for the data mining application

Amassing far more amounts of data is referred to as files exploration. In the beginning, with the entire introduction connected with pc in addition to means for mass digital storage space, we started gathering in addition to holding all kinds of files, relying on the facility connected with PCs to help you examine this particular amalgam connected with data. The actual proliferations connected with database management methods in addition have led in order to latest massive get together connected with all kinds of data.

There are also many other terms, appearing in some articles and documents, carrying a similar or slightly different meaning, such as knowledge mining from databases, knowledge extraction, data archaeology, data dredging, data analysis, etc. By knowledge discovery in databases, interesting knowledge, regularities, or high-level information can be extracted from the relevant sets of data in databases and be investigated from different angles, and large databases thereby serve as rich and reliable sources for knowledge generation and verification. Mining information and knowledge from large databases has been recognized by many researchers as a key research topic in database systems and machine learning and by many industrial companies as an important area with an opportunity of major revenues. The discovered knowledge can be applied to information management, query processing, decision making, process control, and many other applications. Researchers in many different fields, including database systems, knowledge-base systems, artificial intelligence, machine learning, knowledge acquisition, statistics, spatial databases, and data visualization, have shown great interest in data mining.

Furthermore, several emerging applications in information providing services, such as on-line services and World 'Wide Web, also call for various data mining techniques to better understand user behavior, to meliorate the service provided, and to increase the business opportunities.

II. BACKGROUND

Bhat P. et al. [2014] proposed the heuristic based algorithm for hiding the sensitive association rules the algorithm is named as MDSRRC, owner hide sensitive association rule and place transform rules to the server for outsourcing purpose. In this algorithm they are providing an incremental association rule for mining. The proposed algorithm which is Matrix Apriori algorithm based on analysis of two association algorithm named as Apriori algorithm and FP-growth algorithm. The matrix Apriori algorithm which has simple structure similar as a matrices and vectors; the algorithm generates frequent patterns and minimizes the number of sets, as compared to previous algorithm. The matrix algorithm is very simple and has an efficient way to generate association rule than the previous algorithm. For hiding the sensitive information of the database proposed algorithm MDSSRC selects the transactions and items by using certain criteria which transform. As per comparing with the previous algorithm the proposed algorithm is much better in performance which can be concluded with the results of the implementation.

Islam A, et al. proposed improved Frequent Pattern Growth Tree. As finding of association rules from large number of item sets measured as key aspect of data mining. The rising demand of discovering pattern from large data enhances the association rule mining. Then in difficulty occurs in generating the candidate sets.

Jabbar, M. A., et al. Implemented CBARBSN Cluster Based Association Rule Mining Based on Sequence Number in which they proposed a new algorithm which combines the concept of sequence numbers and Clustering. .The entire data base is divided into partitions of equal size, each partition will be called cluster. Each cluster is considered one at a time by loading the first cluster into memory and calculating frequent item sets. Then the second cluster is considered similarly and calculating frequent item sets .

Jain Y. et al. proposed an efficient Association rule hiding algorithm for privacy preserving data mining. This is based on association rule hiding approach of previous algorithms and modifying the database transactions so that the confidence of the association rule can be reduce.

Jain Y. et al. proposed an efficient Association rule hiding algorithm for privacy preserving data mining. This is based on association rule hiding approach of previous algorithms and modifying the database transactions so that the confidence of the association rule can be reduce.

Kaur C. presented a survey of most recent research work. However association rule mining is still in a stage of exploration and development. There are still some essential issues that need to be studied for identifying useful association rules. Some problems for association rule mining are also suggested.

Natarajan et al. presented the privacy problem by considering the privacy and algorithmic requirements simultaneously. The objective of this paper was to implement a algorithm which is a association rule for privacy preserving data mining which would be efficient in providing confidentiality and improve the performance at the time when the database stores and retrieves huge amount of data. This paper compared the performance of proposed algorithm with the two existing algorithms namely ISL and DSR.

Niti Guru et al. proposed a system that uses neural network for prediction of Kidney disease, blood pressure and sugar. A set of 78 records with 13 attributes are used for training and testing. He suggested supervised network for diagnosis of Kidney disease and trained it using back propagation algorithm. On the basis of unknown data is entered by doctor the system will find that unknown data from training data and generate list of possible disease from which patient can suffer.

Shah et al. described association rule hiding approaches and surveyed existing algorithm for association rule hiding. Based on this, comparative analysis of heuristic algorithms described.

Thakur D et al. discussed the basic of PPDM and its different approaches. Subsequently, association rule hiding approaches and metrics for performance comparison of those approaches are discussed. This paper provided an overview of heuristic approaches.

Zhang et al. proposed an oval method to generate association rules that focus on shortest length among template along with minimum support and minimum confidence threshold. This method transforms the multi-valued information system into two-valued information system. Binary relation on attributes is taken from two-valued information and figure outs the topology of attributes based on binary relation.

Ziauddin et al showed a survey of association rule mining since its beginning. Association Rule Mining is one of the key areas of research, receiving rising notice. It is a necessary part of Knowledge Discovery in Databases (KDD). The scale of Association Rule Mining and KDD is very broad. Although it has been come into sight as a new technology but Association Rule Mining is still in a stage of exploration and development.

III. PROPOSED TECHNIQUE

The proposed algorithm named as Critical Relevant Information Description Mechanism (CRIDM) algorithm. It has equally performed or outperformed the existing information discovery algorithms in terms of data classification. The initial design analysis of the algorithm has stated that when the performance parameter of elapsed time would be tested on the latter mentioned four databases it will yield good and acceptable results. An equal or better data representation function values with less elapsed time will be calculated to prove the proposed algorithm better than existing algorithms for large datasets. In future, this algorithm will be tested with an adequate number of datasets and will be compared with the existing information discovery or classification algorithms in the terms of other performance parameters also. Its performance will be also tested and compared with other similar algorithms on the basis of various datasets and more performance parameters. Because the proposed algorithm is proved to be useful for the disease pattern mining, it will be enhanced to perform better than the proposed algorithm by combining it with different algorithms to develop new algorithms using new algorithmic combinations or newly developed algorithms. The system design of proposed research is

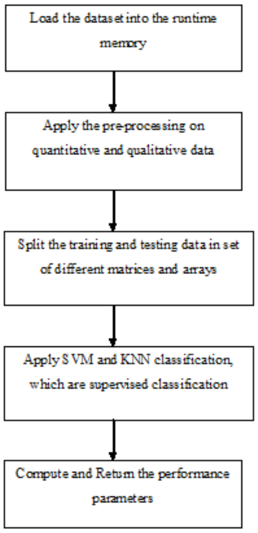
Load the dataset into the runtime memory
Apply the pre-processing on quantitative and qualitative data
Split the training and testing data in set of different matrices and arrays
Apply SVM and KNN classification, which are supervised classification
Compute and Return the performance parameters

Figure 1: The flowchart of system design

The methodology for proposed technique is as follows:

Data Collection and understanding: Dataset were mainly collected from UCI repository and from various hospitals. Patient's data were collected which contains 8 attributes In this, collection of original data is made. Activities are performed in order to get familiar with the data. The data set collected has been entered to the appropriate data mining tool and explored so that the researcher understood the data set properly.

Data Mining Tool Selection: Selection of appropriate data mining tools and techniques depends on the main task of the data mining process. The selected software should be able to provide the required data mining functions and methodologies. The data mining function carried out in this research is classification of patient's dataset, WEKA version 3.6.1010 is used for this purpose.

Classification: Feature vector of training set is fed to learn model to assign a class label to given test samples in the form of an unclassified set. These Feature vectors will be analyzed on the final stage and compare with the feature descriptor vectors obtained from the each data entry in the test and training matrices**.**

In the proposed algorithm, the scaling of the data is being processed before undergoing the Euclidean distance in order to avoid the column dominance. The categorical feature handling must possess the steps to ensure the maximum classification accuracy, which includes missing values followed by numeric labeling & dummy variable creation with one column removal.

$$KCR = \log_{10}\left(\frac{N_{keypoints}}{N_{corners}}\right) \overset{?}{\leq} T_1. \qquad (1)$$

To fit the supervised classification model over the set of the points in the given dataset for the major five stock markets, which includes the data of technology, political, business and sport news in the text file form encoded "UTF-8" encoding. The standard supervised classification equation stated, y = mx + b, is used for the supervised classification model in this price prediction model, where the m denotes the line's or curve's slope, b denotes the y-intercept over the line and x gives the properties of data in the given data. For the best fitting and prediction, we need to utilize the best set of the points around the slope (denoted m) and intercept curve or values (denoted b) with the y-intercept.

The standard error functions are used to compute the error or cost using the supervised classification equations, which takes the input of the data values and return the cost, which defines the predicted values of the price in our model. The conventional to squared distance is computed to ensure the value as positive for the prediction of the price with flexibility. The normal equation is given as the following:

$$\text{Error}_{(\text{m,b})} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + b))^2$$

Where the influential factors (quantity and number of suppliers) have been denoted by m and b respectively, x gives the price data and y denotes the y-intercept to discover the new trend. N represents the total number of the data rows, whereas Error $_{(\text{m,b})}$ gives the result computed by the normal equation.

Gradient Descent is also an error or cost function to predict the future price. The equation of the gradient descent can be given with the following equation:

$$\frac{\partial}{\partial \text{m}} = \frac{2}{N}\sum_{i=1}^{N} -x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial \text{b}} = \frac{2}{N}\sum_{i=1}^{N} -(y_i - (mx_i + b))$$

Where the influential factors (quantity and number of suppliers) have been denoted by m and b respectively, x gives the price data and y denotes the y-intercept to discover the new trend, specifically, in y_i, i index gives index of the different price entities. N represents the total number of the data rows, whereas There are two types of gradient descent, m and b, which is denoted by the (delta / delta * m) and (delta / delta * b) in the above normal equation.

Simple Supervised Classification Simple Supervised Classification: The model for Simple Supervised classification is given by **Y** $=\beta_0 + \beta_1 x + \beta_2 x + \cdots + \beta_n x + \varepsilon$, where

- **Y** is the dependent variable
- X is the independent variable
- **Ɛ** is the random error variable
- $\beta_{\circ}$ is the y-intercept of the line $y = \beta_1 + \beta_0 x$
- β_1 is the slope of the line $y = \beta_1 + \beta_0 x$

In the model above:

Y and X are assumed to be **correlated**, i.e., linearly related, and thus the model function takes the form of a line, $\boldsymbol{Y} = \beta_0 + \beta_1$ X. Although we have discussed the complete algorithm already , which elaborates the overall working of the simple supervised classification classifier for the test of validity of this hypothesis deciding the category of the news data. The simple linear classification revolves around the fitting of the equation with all of the independent variable and coefficients as the equation design. The final result is derived from the list of squared distances, which defines the real-time differences from the training data. The match with the lowest distance decides the class or category of the target news data.

IV. EXPERIMENTAL RESULTS

The proposed model has been designed for the classification of the kidney based health examination data from nearly 400 patients in the given dataset. This data contains the various parameters, which includes blood pressure, age, red blood cell count, white blood cell count, etc. In this thesis, the SVM and KNN classifiers are applied to the dataset in order to obtain the results.

1. A document or an individual entity is broken on the basis of quantitative and qualitative variables.
2. The quantitative and qualitative variables are handled differently in order to create the balanced version of these variables.
3. Finally, both of the feature matrixes, both obtained from quantitative and qualitative, are combined together to create the feature matrix.

Afterwards, the data is divided into training and testing dataset, which is done using the random selection by creating the random number series. The cross validation split works on the 75:25 ratios, which divided the 400 samples into 300 for training and 100 for testing.
The comparison between the classification accuracy, precision, recall, f1 error and statistical errors has been conducted in this section.

Performance metrics

i. **Precision and recall -** Precision and recall are the two metrics that are widely for evaluating performance in text mining, and in text analysis field like information retrieval. These parameters are used for measuring exactness and completeness respectively.

$$\text{Precision}=\frac{\text{True Positive}}{\text{True Positive}+\text{False Positive}} \qquad \text{Eq. (1)}$$

$$\text{Recall}=\frac{\text{True Positive}}{\text{True Positive+False Negative}} \qquad \text{Eq. (2)}$$

ii. **F-measure** -F-Measure is the harmonic mean of precision and recall. The value calculated using F-measure is a balance between precision and recall.

$$\text{F measure }=\frac{2*\text{recall}*\text{precision}}{\text{precision+recall}} \qquad \text{Eq. (3)}$$

iii. **Accuracy-** Accuracy is the common measure for classification performance. Accuracy can be measured as correctly classified **instances** to the total number of **instances**, while error rate uses incorrectly classified instances instead of correctly classified instances.

$$\text{Accuracy}=\frac{\text{True Positive+True Negative}}{\text{True Positive+False Positive+True Negative+False Negative}} \qquad \text{Eq. (4)}$$

The comparison includes the average, standard deviation, median, minimum and maximum values. The comparative analysis is supposed to show the clear analysis, and helps to declare the best algorithm. Hence, the averaging factors play the key roles to distinguish the performance. The following table compares both classification, KNN & SVM on the basis of the true type errors from the statistical errors. The SVM is known to produce the more number of true positive on the average than KNN, although the maximum number of is true negative is higher in case of KNN than SVM.

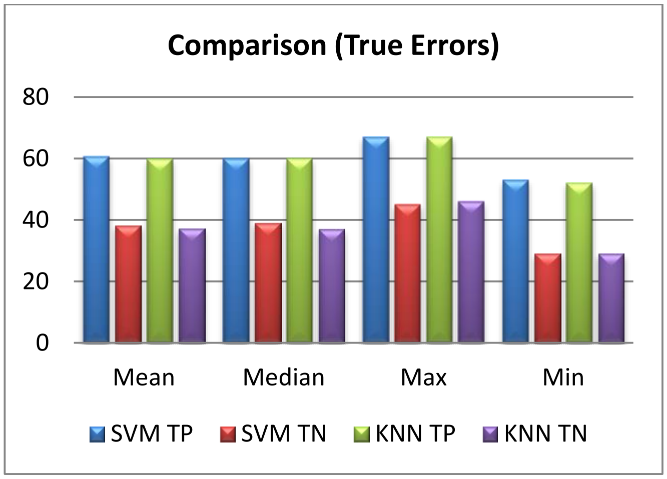

Figure 2: Comparative Analysis of KNN and SVM based on type 1 parameters

The SVM and KNN have been compared on the basis of false type parameters, which include the false positive and false negative parameters. The following table shows the higher performance of SVM on the basis of average false positive cases, which is 1.04 in comparison to 1.88 for KNN. This pattern is identical with lower difference on the basis of false negative.

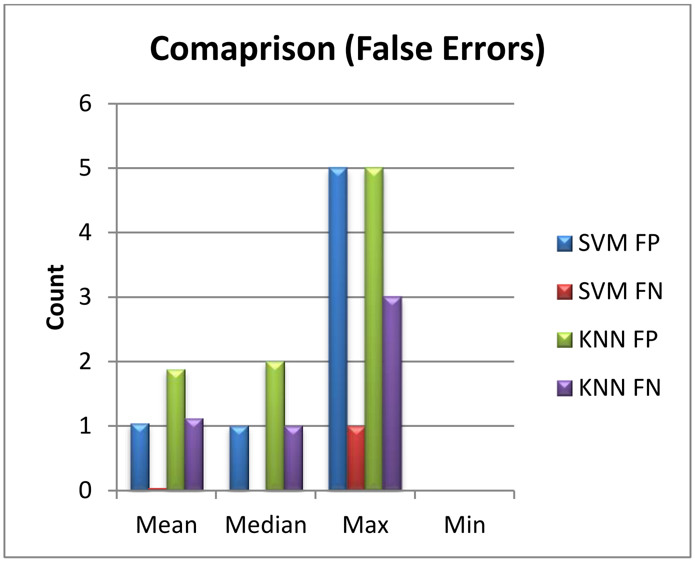

Figure 3: Comparative Analysis of KNN and SVM based on type 2 parameters

The SVM shows the higher performance than KNN on the basis of accuracy, precision and recall based parameters. The following table shows the higher mean for all accuracy, precision and recall than KNN, which clearly signifies the higher performance.

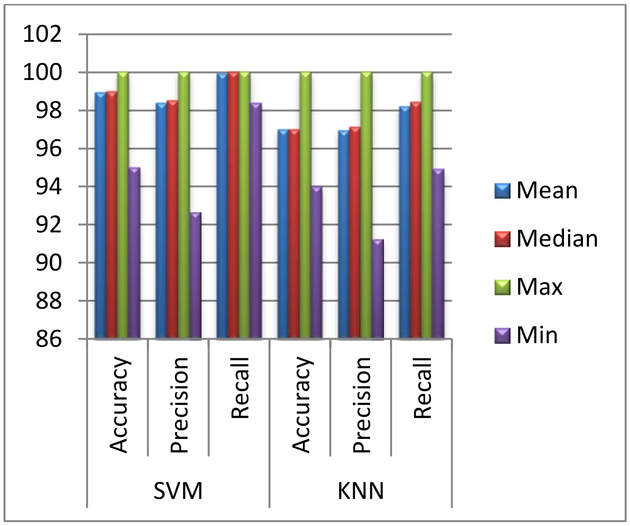

Figure 4: Comparative Analysis of KNN and SVM based on Accuracy, Precision and Recall

V. CONCLUSION

In this research work ,the support vector machine (SVM) and k-nearest neighbor (KNN) based models are used to predict the chronic kidney diseases. The dataset includes 35 features, out of which some of them are continuous and discrete and other all categorical. Different modules are developed to handle the categorical variables in the dataset to avoid the problems related to the column dominance, execution errors, etc. The continuous variables undergo the maximum minimum scaling, which is known to convert the data values to 0-1 scale. In this thesis, SVM outperformed KNN on the basis of precision by (98%) and recall (99%) against 97% (precision) and 98% (recall)**.**

In the future, the use data mining technique on any other medical condition .Here we perform classification In future, the deep learning classification can be utilized to improve the overall classification performance. Also, the optimization algorithms can be used to create the more balanced and advanced feature descriptors to obtain the higher accuracy.

REFERENCES

[1]. Asha Rajkumar, G. Sophia Reena, ―Diagnosis of Kidney disease Using Datamining Algorithm‖; Global Journal of Computer Science and Technology, Page 38 Vol. 10 Issue 10 Ver. 1.0 September, 2010.

[2]. Bhatla, Nidhi, and Kiran Jyoti. "An analysis of heart disease prediction using different data mining techniques." *International Journal of Engineering* 1, no. 8 (2012): 1-4.

[3]. Chowdhury, Dilip Roy, Mridula Chatterjee, and R. K. Samanta. "An artificial neural network model for neonatal disease diagnosis." *International Journal of Artificial Intelligence and Expert Systems (IJAE)* 2, no. 3 (2011): 96-106.

[4]. Dhutraj N.,Sasane S.,Kshirsagar V., “Hiding Sensitive Association Rule for Privacy Preservation”, Institute of Electrical and Electronics Engineers Transactions on knowledge and data engineering,2013 .

[5]. Guru, Niti, Anil Dahiya, and Navin Rajpal. "Decision support system for kidney disease diagnosis using neural network." *Delhi Business Review* 8, no. 1 (2007): 99-101.

[6]. Jabbar, M. A., Priti Chandra, and B. L. Deekshatulu. "Cluster based association rule mining for heart attack prediction." *Journal of Theoretical and Applied Information Technology* 32, no. 2 (2011): 197-201.

[7]. Jadav K.,Vania J., Patel D. (2013), “A Survey on Association Rule Hiding Methods”, International Journal of Computer Applications, Vol. 82 (13), pp-20-25.

[8]. Jyoti Soni, Sunita Soni et al., ―Predictive Data Mining for Medical Diagnosis: An Overview of Kidney disease Prediction‖; International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[9]. Kaur C., “Association Rule Mining using Apriori Algorithm: A Survey”, International Journal of Advanced Research in Computer Engineering & Technology, Vol. 2(6), pp-893-900 ,2013.

[10]. Natarajan R.,Sugumar R., Mahendran M.,Anbazhagan K. , “Design and Implement an Association Rule hiding Algorithm for Privacy Preserving Data Mining”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 1(7), pp.486-492, 2012.

[11]. Niti Guru, Anil Dahiya, Navin Rajpal, Decision Support System for Kidney disease Diagnosis Using Neural Network, Delhi Business Review, Vol. 8, No. 1, January-June 2007.

[12]. Palaniappan, Sellappan, and Rafiah Awang. "Intelligent heart disease prediction system using data mining techniques." In *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*, pp. 108-115. IEEE, 2008.

[13]. Patil, Shantakumar B., and Y. S. Kumaraswamy. "Intelligent and effective heart attack prediction system using data mining and artificial neural network." *European Journal of Scientific Research* 31, no. 4 (2009): 642-656.

[14]. Rani, K. Usha. "Analysis of heart diseases dataset using neural network approach." *arXiv preprint arXiv:1110.2626*(2011).

[15]. Rupa G. Mehta, Dipti P. Rana, Mukesh A. Zaveri, ―A Novel Fuzzy Based Classification for Data Mining using Fuzzy Discretization‖; World Congress on Computer Science and Information Engineering, 2009.

[16]. Shantakumar B.Patil, Y.S.Kumaraswamy, ―Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network‖; European Journal of Scientific Research, ISSN 1450-216X Vol.31 No.4, 2009.