International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

A Classification Approach for instant Medical Assistance to Health Seekers

Pritha Tikariha*, Prashant Richhariya
Department of Computer Science, CSIT, Durg, Chhatishgarh, India

ABSTRACT

Data mining approach is applied in numerous numbers of fields for predicting and forecasting the events. However, in healthcare sectors, due to lack of faith in prediction method people hesitate to utilize data mining technique for health issues. People post their health related queries and get reply from the experts in many online healthcare applications. However, health seekers do not get instant assistance there; they need to wait for the experts for their opinion. Many data is accumulated in repository of such application. Using data mining techniques, useful information can be extracted from such repository, which can help health seekers to get instant assistance for their health related issues. These paper presents analysis on some data mining technique particularly in disease dataset. Three classification algorithms i.e. KNN, SVM, Naïve Bayes are applied on three disease dataset, to analyse the performance of the classifiers. The predictive rate is evaluated using four evaluation parameters i.e. Accuracy, precision, recall and f_measure. The experiment is performed in Matlab tool shows that Naive Bayes outperforms as compared to rest of the classifiers.

Keywords : Arduino, Wi-Fi (ESP 8266), Load cell, Database System

I. INTRODUCTION

In health care domain, data can be very valuable. These data can be mined and converted into useful information. Medical data mining provides a way to explore the hidden relationship present in the data set of the medical realm. This relationship can be used for the diagnosis of many diseases. However, these medical data sets are very huge and obscure. These dataset have to be structure and assimilated to form a medical information system. Medical data mining provides a way to achieve these things.

Patients generally have different medical attribute as a result they have heterogeneous medical requirement. These heterogeneous attribute needs to be classified accordingly and transformed into homogeneous groups. For converting it into homogeneous groups, these dataset require detailed, effective and efficient classification algorithms. Homogeneity brings the benefits of increased certainty in clinical diagnosis, predicting individual patient needs and resource utilization.

In these research work three classification algorithms is applied on different disease dataset, to determine the prediction rate of the classifier. KNN, SVM and Naïve Bayes algorithm are analysed using for evaluation parameters accuracy, precision, recall and f-measure.

The remaining paper is organized in the following way: Section II reviews the previous work done on different dataset and classifiers. Section III describes the methodology and the dataset. Section IV includes the experimental setup results and finally section V has conclusion of the research work and the future scope.

II. LITERATURE SURVEY

Classification algorithms are generally very useful for medicinal issues, especially when applied for the diagnosis purpose [4]. Many machine learning algorithms are applied in the medical domain in the course of recent decades [5] [6] [7] [8] [9] [10].A large portion of these applications are specific and include machine learning procedure like using data mining for diagnosis purpose [11], applied neural network rule for the prediction of breast cancer [10]. Data mining has

been effectively functional in many medical fields. [12] applied KDD for the diagnosis of cardiac SPECT. [13] worked with data mining approach for the assessment of haemodialysis process. [14] utilized data mining for the predicting survivability of kidney dialysis. [15] have presented a survey of recent work done as such far in prediction of cancer diseases. In [16] proposed a system for diagnosis and prediction of CKD based on predictive mining. [17] have applied Artificial Neural Networks, Decision tree and Logical Regression .They have compared the performance of these data mining techniques. These techniques ids tested on the data collected from various dialysis centres.

III. METHODOLOGY

In this research work, three classification techniques are applied on three data set to analyse the prediction rate of the classifiers. The classifier applied is KNN, SVM and Naïve Bayes. These classifiers are applied on the datasets to predict whether disease is present or not. The performance of the each classifier is evaluated based on accuracy, precision, recall and F-measure.

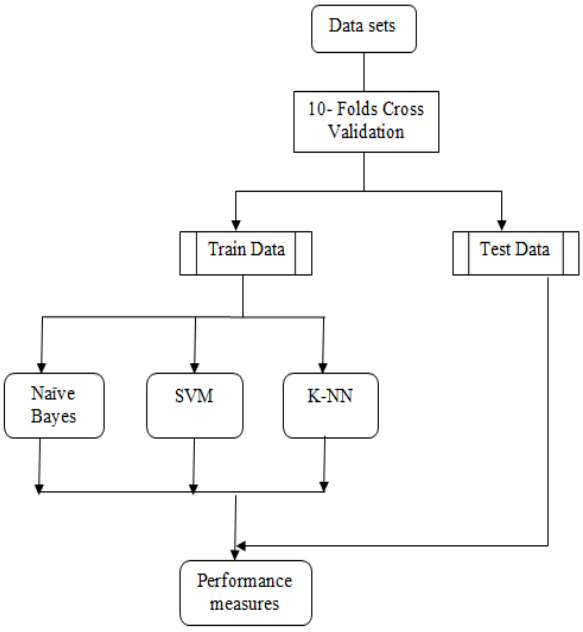

Figure 1. Workflow of predictive data mining

The working of the architecture is as: The dataset is extracted from different repository. Each dataset is divided into test data and train data using 10 fold cross validation different classification algorithm is trained is using train data and the evaluation parameters are calculated. Result of the evaluation will determine the best classifier algorithm for this dataset by comparing the evaluation parameters.

A. Dataset and Pre-processing

The dataset is collected from the UCI repository as an input to the Weka tool for predicting the CKD and notCKD. The data is collected from Apollo Hospital in Tamil Nadu in July 2015. Dataset contains 400 instances that are already classified as CKD and notCKD. There are 250 instances classified as CKD and 150 as notCKD. It has 25 nominal and numeric attributes such as age, blood pressure, sugar, potassium, haemoglobin etc. Out of 25 attributes, last one is the class that can be either CKD or notCKD [1]. The dataset is found to have lots of missing values. Missing values from the dataset is removed in the Weka tool using unsupervised attribute filter. Replace missing values replaces the nominal and numeric missing values in a dataset by taking mean and modes from the training data. [3]. Once the missing values are removed, all three classification algorithms are applied on the pre-processed data.

The dataset is collected from keel dataset [2]. This dataset contains total of 10 attributes some of which are linear and some are nominal. All the attributes are first converted into nominal values. Out of 10 attributes, 9 are used to predict the breast cancer and the last attribute is class which can either be recurrence events or no recurrence events. There are 286 instances out of which 201 instances belong to no recurrence events and 85 instances belong to recurrence event. There are no missing values in the dataset.

Heart disease dataset is again collected from keel dataset. The dataset is gathered from the Cleveland Clinic Foundation, Hungarian Institute of Cardiology, Budapest, V.A. Medical Centre, Long Beach, CA and University Hospital, Zurich, Switzerland [1]. There are 13 attributes which is used to predict whether heart disease is present or not and the last attribute is class. All the attributes are nominal values. There is no missing value found in the dataset. Total of 270 instances are present. Out of which 150 instances belong to the class that does not have heart disease and 120 belong to the class that shows the presence of heart disease.

B. Evaluation Parameters

When we use a classifier model for a problem, we usually want to look at the accuracy of that model as the number of correct predictions from all predictions made. This is the classification accuracy. When we

need to decide whether it is a good enough model to solve the problem, accuracy is not the only metric for evaluating the effectiveness of a classifier. Some other useful metrics are precision, recall and f-measure. These metrics can provide much greater insight into the performance characteristics of a classifier.

- **Accuracy:** Accuracy (Acc) is defined as the number of predictions which are correct.

$$\text{Acc} = \frac{(\text{Tp} + \text{Tn})}{(\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn})}$$

- **Precision:** Precision (pre) is defined as the probability by which the randomly selected instance is relevant.

$$\text{Pre} = \frac{\text{Tp}}{(\text{Tp} + \text{Fp})}$$

- **Recall:** Recall (Re) is defined as the probability that the randomly selected instance is relevant in the search.

$$\text{Re} = \frac{\text{Tp}}{(\text{Tp} + \text{Fn})}$$

- **F-Measure:** F-Measure (F_{mean}) is calculated as the harmonic mean of precision and recall.

$$\text{F}_{\text{mean}} = \frac{2 * (\text{Pre} * \text{Re})}{(\text{Pre} + \text{Re})}$$

Where,

Tp: the number of true positives.

Fn: the number of false negative.

Fp: the number of false positives.

Tn: the number of true negatives.

IV. EXPERIMENTAL RESULTS

This work is carried on Matlab tool. It has many inbuilt algorithm but at the same time it allows the researchers to implement their own algorithms. The pre-processed data is taken as input to the Matlab tool and classification algorithms are applied on it. The performance of the classifiers is evaluated on the basis of various evaluation parameters.

A. Result of Chronic Kidney Disease Dataset

From the Table 1 it is observed that Naïve Bayes performs best as compared to SVM and KNN with the highest accuracy of 0.975.

TABLE I

RESULT FOR CHRONIC KIDNEY DISEASE DATASET

	Accuracy	Precision	Recall	F-Measure
KNN	0.9625	1	0.956	0.9772
SVM	0.9605	0.952	0.942	0.971
Naïve Bayes	0.9725	0.965	0.992	0.9782

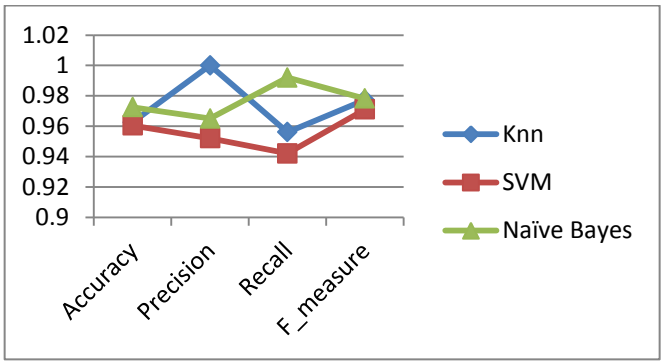

Figure 2. Graph for Chronic Kidney Disease

B. Result of Heart Disease Dataset

From Table 2 and Fig 3, we can conclude that for the heart disease dataset again Naive Bayes outperforms as compared to KNN and SVM. Accuracy for Naïve Bayes classifier is 0.829.

TABLE 2

RESULT FOR HEART DISEASE DATASET

	Accuracy	Precision	Recall	F-Measure
KNN	0.807	0.813	0.86	0.833
SVM	0.811	0.794	0.8933	0.8404
Naïve Bayes	0.829	0.839	0.86	0.847

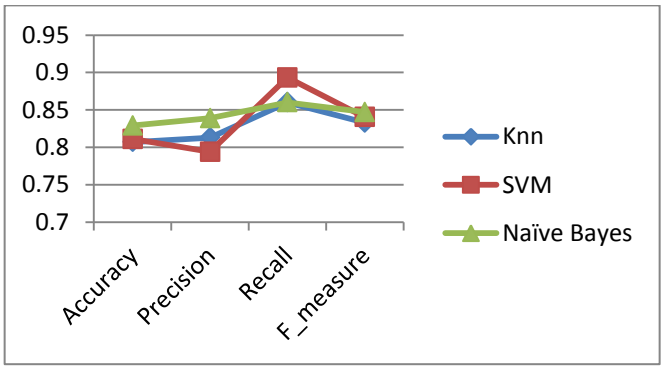

Figure 3. Graph for Heart Disease

C. Result of Breast Cancer Dataset

From Table 3 shows that accuracy of Naïve Bayes is 0.74 which is best among all the classifier.

TABLE 3
RESULT FOR HEART DISEASE DATASET

	Accuracy	Precision	Recall	F-Measure
KNN	0.686	0.475	0.275	0.331
SVM	0.735	0.433	0.309	0.33
Naïve Bayes	0.74	0.703	0.254	0.36

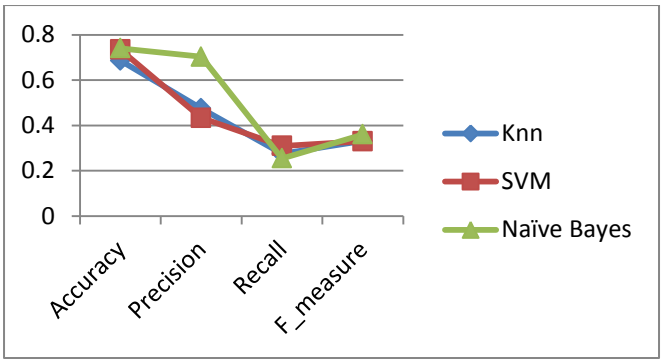

Figure 4. Graph for Heart Disease

When KNN, SVM and Naïve Bayes are applied on three medical dataset, rate of prediction is best for Naïve Bayes with the highest accuracy as compared to the rest of two classifiers.

V. CONCLUSION & FUTURE WORK

Three classification techniques i.e. KNN, SVM and Naïve Bayes algorithms are used here. These three classifiers are applied on three different dataset. Chronic kidney disease, Heart disease and breast cancer dataset are utilized for the purpose of analyzing the classifiers. When the datasets are applied to the classifiers, Naïve Bayes classifier outperforms from the rest. But still there was a chance to improve the accuracy of the classifier. There are many ensemble techniques which can boost the performance of the classifiers. These classifiers can be fused with the ensemble techniques and can be utilized to give more accurate results.

VI. REFERENCES

[1] UCIrepository:archive.ics.uci.edu/ml/datasets/chronic_kidney_disease

[2] Keel.es : http://sci2s.ugr.es/keel/category.php?cat=clas

[3] WEKATool:www.cs.waikato.ac.nz/ml/weka/downloading.html.

[4] Freitus, A. A. (2002). A survey of evolutionary algorithms for data mining and knowledge discovery. In A. Ghosh & S. Tsutsui (Eds.), Advances in evolutionary computation. Berlin: Springer.

[5] Becerra-Fernandez, I. (2000). The role of artificial intelligence technologies in the implementation of People-Finder knowledge management systems. Knowledge-based Systems, 13, 315–320.

[6] Evans, C. D. (1999). A case-based assistant for diagnosis and analysis of dysmorphic syndromes. Med Inform, 20, 121–131.

[7] Kukar, M., Kononenko, I., & Silvester, T. (1996). Machine learning in prognosis of the femoral neck fracture recovery. Artificial Intelligence in Medicine, 8, 431–451.

[8] Sacha, J. P., & Cios, K. J. (2000). Issues of in automating cardiac SPECT diagnosis. IEEE Engineering in Medicine and Biology Magazine, 19(4),78–88.

[9] Schmidt, R., & Gierl, L. (2001). Case-based reasoning for antibiotics therapy advice: an investigation of retrieval algorithms and prototypes. Artificial Intelligence in Medicine, 23, 171–186.

[10] Setiono, R. (1996). Extracting rules from pruned neural networks for breast cancer diagnosis. Artificial Intelligence in Medicine, 8(1), 37–54.

[11] Alonso, F., Caraca-Valente, J. P., Gonzalez, A. L., & Monte, C. (2002).Combining expert knowledge in a medical diagnosis domain. Expert Systems with Applications, 23, 367–375.

[12] Kurgan, L. A., Cios, K. J., Tadeusiewicz, R., Ogiela, M., & Goodenday, L. S. (2001). Knowledge discovery approach to automated cardiac SPECT diagnosis. Artificial Intelligence in Medicine, 23(2), 149–169.

[13] Bellazzi, R., Larizza, C., Magni, P., & Bellazzi, R. (2005). Temporal data mining for the quality assessment of hemodialysis services. Artificial Intelligence in Medicine, 34(1), 25–39.

[14] Kusiak, A., Dixon, B., & Shah, S. (2005). Predicting survival time for kidney dialysis patients: a data mining approach. Computers in Biology and Medicine, 35(4), 311–327.

[15] Konstantina Kourou et.al, "Machine learning applications in cancer prognosis and prediction" Computational and structural biotechnology Journal, Elsevier.

[16] P.Swathi Baby, T. Panduranga Vital,"Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms" IJERT.

[17] K.R.Lakshmi1, Y.Nagesh2 and M.VeeraKrishna3, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability", IJAET.