Analysis of various data mining classification techniques to predict diabetes mellitus

T.Nithyapriya, S.Dhinakaran
M.Phil. Scholar, Assistant Professor
Department of Computer Science,
Rathinam College of Arts and Science, Coimbatore, India

Abstract—Data mining approach helps to diagnose patient’s diseases. Diabetes Mellitus is a chronic disease to affect various organs of the human body. Early prediction can save human life and can take control over the diseases. This research work explores the early prediction of diabetes using various data mining techniques. The real time diabetic based dataset has taken with 203 instances for training data set and 52 instances for test data set to determine the accuracy of the Naïve Bayes, SVM and J48 classification techniques in prediction. The analysis proves that SVM Classifier provide the highest accuracy than other techniques.

Index Terms—Data mining, Diabetes, Prediction, accuracy, classification

I. INTRODUCTION

Data mining take holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Some experts believe the opportunities to get better care and reduce costs simultaneously could apply to as much as 30% of overall healthcare spending. This could be a win/win overall. But due to the complexity of healthcare and a slower rate of technology adoption, our industry lags behind these others in implementing effective data mining and analytic strategies.

Data mining has become an essential methodology for computing applications in medical informatics. Progress in data mining applications and its implications are manifested in the areas of information management in healthcare organizations, health informatics, epidemiology, patient care and monitoring systems, assistive technology, large-scale image analysis to information extraction and automatic identification of unknown classes. Various algorithms associated with data mining have significantly helped to understand medical data more clearly, by distinguishing pathological data from normal data, for supporting decision-making as well as visualization and identification of unseen complex relationships between diagnostic features of different patient groups.

II. DATA MINING IN DIABETIC MELLITUS

Diabetes mellitus, or simply diabetes, is a set of related diseases in which the body cannot regulate the amount of sugar in the blood [4]. It is a group of metabolic diseases in which a person has high blood sugar, either because the body does not produce enough insulin, or because cells do not respond to the insulin that is produced. This high blood sugar make the classical symptoms of polyuria, polydipsia and polyphagia [5]. There are three main types of diabetes mellitus (DM). Type 1 DM results from the body's failure to produce insulin, and presently requires the person to inject insulin or wear an insulin pump. This form was before referred to as "insulin dependent diabetes mellitus" (IDDM) or "juvenile diabetes". Type 2 DM results from insulin resistance, a condition in which cells fail to use insulin properly, sometimes combined with an absolute insulin deficiency. This form was previously referred to as noninsulin dependent diabetes mellitus (NIDDM) or "adult-onset diabetes". The third main form, gestational diabetes occurs when pregnant women without a previous diagnosis of diabetes develop a high blood glucose level. It may lead development of type 2 DM.

Data Mining [6] refers to extract or mining knowledge from huge amounts of data. The aim of data mining is to make sense of huge amounts of mostly unsupervised data, in some domain. Classification [1] maps data into predefined groups. It is often referred to as supervised learning as the classes are determined prior to examining the data. Classification Algorithms usually require that the classes be defined based on the data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to class. Pattern Recognition is a type of classification where an input pattern is classified into one of the several classes based on its similarity to these predefined classes. Knowledge Discovery in Databases (KDD) is the process of finding useful information and patterns in data which involves Selection, Pre-processing, Transformation, Data Mining and Evaluation.

Diabetic mellitus in India

In 2000, India (31.7 million) became topped the world with the highest number of people with diabetes mellitus followed by China (20.8 million) and United States (17.7 million) with diabetes mellitus'. According to Wild et al [7] the prevalence of diabetes is predicted to twice globally from 171 million in 2000 to 366 million in 2030 with a maximum increase in India. In 20130, 79.4 million individuals in India will be affected by diabetic mellitus, while China (42.3 million) and the United States (30.3 million) will also increases in those affected by the disease.
Nowadays, India became the diabetes capital of the world with as many as 50 million people suffering from type-2 diabetes, India has a challenge to face. However, medical specialists feel that timely detection and right management can go a long way in helping patients lead a normal life.

India having the highest number of diabetic patients in the world, the sugar disease is posing an enormous health problem to our country today including Pressure, time taken to heal the wounds, tiredness, blurred vision, etc., Often known as the diabetes capital of the world, India has been observing an alarming rise in incidence of diabetes according to the International Journal of Diabetes in Developing Countries. According to a World Health Organization’s diabetes fact sheet, an estimated 34 lakhs deaths are caused due to high blood sugar.

**Purpose of Study:**

This study has aims to implement several prediction based classification techniques in data mining to assist medical institutions, medical research centres and labs with predicting the people’s diabetic mellitus status. If the persons are predicted to have a chance to affected by the diabetic mellitus, then extra efforts can be made to improve their health conditions and allows to suggest the necessary steps to be taken to protect their health from diabetic mellitus.

### III. MEDICAL DATA MINING USING CLASSIFICATION

Medical Data mining refers to extracting or "mining" knowledge from large amounts of medical data. Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. Currently, the data are stored in diabetic database, these database contain the useful information to predict diabetic mellitus. The most useful data mining techniques in medical database is classification.

We have Figure 3.1 that represents working methodology based on the framework. It is important to have a working methodology to govern our work before applying data mining techniques. The work methodology begins with problem definition, data collection and data preprocessing that includes data selection and data transformation and it precedes with data mining classification techniques with pruning which leads to discovering knowledge that is benefit to us.

**Figure 3.1. Data mining work methodology**

- **Data Set**
  - Data collection questionnaire consists of 17 questions with sub-questions such as Name, Age, Weight, Physical activity, Urination, Water consumption, Diet, Systolic blood pressure, Hyper tension, Tiredness, Blurred vision, Wound healing, Sleepy/drowsy, Sudden weight loss, Heredity, Glucose level and Diabetic Mellitus presented.
  - Total size of the data set is 255 with 17 attributes. Collected all details are stored in Excel spreadsheet file (xls) format. It is used to predict the diabetic mellitus in the test data set using classification techniques.

**Classification methods used**

In this research work the following classification methods are used to predict the diabetic mellitus and also analyse the performance of these classification techniques in the diabetic data set

- Naïve Bayes
- Support Vector Machine
- J48
Attribute Selection

In those fields were chosen which were requisite for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database. All the predictor and response variables which were derived from the database are given in Table 3.1 for reference.

Table 3.1 Selected attributes

<table>
<thead>
<tr>
<th>Variables</th>
<th>Description</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age in years</td>
<td>{1 to 100}</td>
</tr>
<tr>
<td>Weight</td>
<td>Weight in Kg’s</td>
<td>{5 to 120}</td>
</tr>
<tr>
<td>Physical activity</td>
<td>Physical activity in minutes</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Urination</td>
<td>Number of times urination in a day</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Water consumption</td>
<td>Water consumption in litres</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Excessive Hunger</td>
<td>Excessive hunger in day time</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>Enter value of blood pressure in “mmHg”</td>
<td>{50 to 200}</td>
</tr>
<tr>
<td>Hyper tension</td>
<td>Person with hyper tension</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Tiredness</td>
<td>Feel tiredness</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Blurred vision</td>
<td>Have blurred vision</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Wound healing</td>
<td>Wound Healing quickly</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Sleepy/drowsy</td>
<td>Always feel sleepy/drowsy</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Sudden weight loss</td>
<td>Observed sudden weight loss</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Heredity</td>
<td>Elders found with diabetes</td>
<td>[Yes, No]</td>
</tr>
<tr>
<td>Glucose level</td>
<td>Level of glucose in blood(No)</td>
<td>{50 to 400}</td>
</tr>
<tr>
<td>Diabetic</td>
<td>Diabetic Present</td>
<td>[Yes, No]</td>
</tr>
</tbody>
</table>

Fig. 3.2. Training Data Set
IV. CLASSIFICATION RESULTS

4.3.1. NAIVE BAYES CLASSIFICATION

Naive Bayes doesn't select any important features. The result of the training of a Naive Bayes classifier is the mean and variance for every feature. The classification of new samples into 'Yes' or 'No' is based on whether the values of features of the sample match best to the mean and variance of the trained features for either 'Yes' or 'No' for the diabetic class variable.

In this test data 97.5369% of diabetic training data instances are correctly classified and remaining 2.4631% of diabetic instances are incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy. So this data set consists of 97.5% accurate instances. The raw numbers are shown in the confusion matrix, with a and b representing the class labels. Here there were 203 instances, so the percentages and raw numbers add up, \( a + b = 74+124 = 198 \), \( a + b = 3+2 = 5 \).

Kappa is a chance-corrected measure of agreement between the classifications and the true classes.

**TP Rate**: rate of true positives (instances correctly classified as a given class). Weighted average TP Rate of this data set is 0.975.

**FP Rate**: rate of false positives (instances falsely classified as a given class). Weighted average FP Rate of this data set is 0.025.

**Precision**: proportion of instances that are truly of a class divided by the total instances classified as that class. Weighted average Precision value of this data set is 0.975.

**Recall**: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). Weighted average Recall of this data set is 0.975.
F-Measure: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall). Weighted F-Measure value is 0.975.

Fig.4.2. Test data set with Predicted Result and Predicted Margin

The correctly and incorrectly classified instances show the percentage of training instances that were correctly and incorrectly classified. In this test data 99.5074% of training data instances are correctly classified and remaining 0.4926% of instances are incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy. So this data set consists of 99.5% accurate instances. The raw numbers are shown in the confusion matrix, with a and b representing the class labels. Here there were 203 instances, so the percentages and raw numbers add up, aa + bb = 76+126 = 202, ab + ba = 1+0 = 1.

Kappa is a chance-corrected measure of agreement between the classifications and the true classes. It's calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement. A value greater than 0 means that this classifier is doing better than chance.

TP Rate: rate of true positives (instances correctly classified as a given class). Weighted average TP Rate of this data set is 0.995.

FP Rate: rate of false positives (instances falsely classified as a given class). Weighted average FP Rate of this data set is 0.003.

Precision: proportion of instances that are truly of a class divided by the total instances classified as that class. Weighted average Precision value of this data set is 0.995.
Recall: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). Weighted average Recall of this data set is 0.995.

F-Measure: A combined measure for precision and recall calculated as $2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall})$. Weighted F-Measure value is 0.995.

In this diabetic test data 98.5222% of training data instances are correctly classified and remaining 1.4778% of instances are incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy. So this data set consists of 98.5% accurate instances. The raw numbers are shown in the confusion matrix, with a and b representing the class labels. Here there were 203 instances, so the percentages and raw numbers add up, $a + b = 74 + 126 = 200$, $ab + ba = 1 + 2 = 3$.

TP Rate: rate of true positives (instances correctly classified as a given class). Weighted average TP Rate of this data set is 0.985.

FP Rate: rate of false positives (instances falsely classified as a given class). Weighted average FP Rate of this data set is 0.019.

Precision: proportion of instances that are truly of a class divided by the total instances classified as that class. Weighted average Precision value of this data set is 0.985.

Recall: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). Weighted average Recall of this data set is 0.985.
**F-Measure:** A combined measure for precision and recall calculated as $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$. Weighted F-Measure value is 0.985.

Fig. 4. Test data set with Predicted Result and Predicted Margin by J48

---

**RESULTS AND DISCUSSIONS**

5.1. **COMPARISON OF CLASSIFICATION ALGORITHMS BASED ON CLASSIFIED INSTANCE**

<table>
<thead>
<tr>
<th>CLASSIFICATION ALGORITHM</th>
<th>CORRECTLY CLASSIFIED INSTANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAÏVE BAYES</td>
<td>198</td>
</tr>
<tr>
<td>SVM</td>
<td>202</td>
</tr>
<tr>
<td>J48</td>
<td>200</td>
</tr>
</tbody>
</table>

The above table reveals that out of 203 instances, 202 instances are correctly classified by the Support vector machine, 200 instances are correctly classified by the J48 classifier and 198 instances are correctly classified by the Naïve Bayes.

Support Vector Machine produced highest accuracy (99.50%) in the classification of diabetic data set. J48 classifier produced 98.52% accuracy and Naïve Bayes Classifier Produced 97.54% accuracy in the classification of diabetic data set.

Chart 5.1. Correctly classified instances of various classifiers

---

**Prediction Result for Correctly Classified Instance**

<table>
<thead>
<tr>
<th>Prediction Result</th>
<th>100</th>
<th>99</th>
<th>98</th>
<th>97</th>
<th>96</th>
<th>95</th>
<th>94</th>
<th>93</th>
<th>92</th>
<th>91</th>
<th>90</th>
<th>89</th>
<th>88</th>
<th>87</th>
<th>86</th>
<th>85</th>
<th>84</th>
<th>83</th>
<th>82</th>
<th>81</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>97.5369</td>
<td>99.5074</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98.522</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.2. F-MEASURE VALUES FOR EACH CLASSIFICATION MODEL

<table>
<thead>
<tr>
<th>CLASSIFICATION MODEL</th>
<th>F-MEASURE VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAÏVE BAYES</td>
<td>0.975</td>
</tr>
<tr>
<td>SVM</td>
<td>0.995</td>
</tr>
<tr>
<td>J48</td>
<td>0.985</td>
</tr>
</tbody>
</table>

The above table reveals that the classification process in the 203 instances, F-Measure value of Support Vector Machine is 0.995, J48 classifier’s F-Measure value is 0.985 and Naïve Bayes classifier’s F-Measure value is 0.975.

5.3. COMPARISON MODELS

To calculate the performance of the various classification models, Correct Classified Rate (CCR), Recall Rate (RR), and F-measure to be used in document or data set classification criteria. The CCR is the rate of correct prediction, and Recall Rate is the ratio actually hit accurate predictions. And F-measure means the combinational mean of CCR and RR, and this is convenient expression method to compare models. The accuracy (AC) is the proportion of the total number of predictions that were correct. The recall or true positive rate (TP) is the proportion of positive cases that were correctly identified. The false positive rate (FP) is the proportion of negatives cases that were incorrectly classified as positive.

Support Vector Machine classification model has the best F-measure value (0.995), Correctly classified rate (99.5%), best Recall rate (0.995) and best correctly classified instances (202) model showed the value compared to the other models such as Naïve Bayes and J48 decision tree in data classification of the diabetic data set with 203 instances. J48 classification model is closely followed to the Support vector machine model.

VI. CONCLUSION AND FUTURE ENHANCEMENT

Data Classification is an important application area in prediction mining in the medical data sets why because classifying millions of patient’s records manually is an expensive and time consuming task. Therefore, automatic classifier is constructed using pre classified sample diabetic data set whose accuracy and time efficiency is much better than manual classification and prediction. Identifying efficient patterns also plays major role in text classification. Data mining classification techniques need to be designed to effectively manage large numbers of elements with varying frequencies. Almost all the known techniques for classification such as decision trees rules, Bayes methods and SVM classifiers have been used to the case of diabetic data.

In this research work, training and test diabetic data sets are used to predict the diabetic mellitus using various classification techniques. And we compared those data by applying the material to the conventional techniques of Bayesian statistical classification, J48 Decision tree and SVM to form a prediction model. The SVM model shows better performance than J48 and Naïve Bayes classification models. Future works may also include hybrid classification models by combining some of the data mining techniques such as attribute selection and clustering.

REFERENCES