CHAPTER 7
CASE STUDY ON UCI DATASET

7.1 INTRODUCTION

Diabetes disease has become more prevalent in recent years, prompting scholars to devote more attention to its risk factors. Early diagnosis and treatment is one of the best approaches to reduce the disease death rate. For such reasons, Knowledge Discovery techniques may be a practical and effective solution for generating and proving new hypothesis, mining and generalizing new medical knowledge directly from pertinent real examples.

In this chapter, an approach to apply a real time dataset on our new model PRMOTE is presented. Section 7.2 describes the dataset used in the example for construction of the model. Section 7.3 discusses the preparation of dataset for C4.5 and PRMOTE model. Section 7.4 gives the parameters on which the evaluation of the results will be done. Section 7.5 compares the experimental results and discusses the obtained results. In Section 7.6 a summary of the chapter and a conclusion is provided.

7.2 The Data Set Used

7.2.1 Collection of the Dataset

In this case study, the goal is to diagnosis whether a patient has diabetes or not. The present data analysis was carried out on a data collection of diabetes disease cases for female taken from UCI repository. This UCI dataset was contributed by V. Sigillito. This data set is used to diagnose whether a
patient has diabetes or not. All patients in the data set are females, at least 21 years old, and of Pima Indian heritage living near Phoenix, Arizona, USA. Number of instances in the data set is 768, number of attributes is 8, and number of classes is 2 (0=health, 1=tested positive for diabetes). There are 500 of the examples which are healthy and the others are tested positive for diabetes. There are no missing attribute values.

Table 7.1: Attribute information of the pima diabetes dataset taken from UCI repository for the purpose of study.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preg</td>
<td>Number of times pregnant</td>
</tr>
<tr>
<td>Plas</td>
<td>Plasma glucose concentration a 2 hours in an oral glucose tolerance test</td>
</tr>
<tr>
<td>Pres</td>
<td>Diastolic blood pressure (mm Hg)</td>
</tr>
<tr>
<td>Skin</td>
<td>Triceps skin fold thickness (mm)</td>
</tr>
<tr>
<td>Insu</td>
<td>2-Hour serum insulin (mu U/ml)</td>
</tr>
<tr>
<td>Mass</td>
<td>Body mass index (weight in kg/(height in m)^2)</td>
</tr>
<tr>
<td>Pedi</td>
<td>Diabetes pedigree function</td>
</tr>
<tr>
<td>Age</td>
<td>Age (years)</td>
</tr>
<tr>
<td>Class</td>
<td>Class variable (0 or 1)</td>
</tr>
</tbody>
</table>
7.2.2 Dividing the Data Set

We run our new algorithm with 10 runs of 10 fold cross validation. In 768 instances we have taken 691 cases as training set and 77 cases as test set.

7.3 Construction of the model

7.3.1 Original Dataset used by the C4.5 decision tree

We have used the original dataset for construction of decision tree by the C4.5 decision tree. Figure 6.1 shows the distribution of positive and negative instances of diabetes cases. The red marks represent positive tested instances and blue marks represent negative tested instances.

![Figure 6.1: The original distribution of pima diabetes data set used by C4.5 algorithm for generating CIL measure.](image)

Figure 7.1: The original distribution of pima diabetes data set used by C4.5 algorithm for generating CIL measure.
7.3.2 Improved Pima diabetes dataset using PRMOTE

We have improved the same dataset by using our new PRMOTE method. Figure 7.2 shows the misclassified, borderline, noisy and weak instances to be removed from the majority subset before resampling. Figure 6.3 shows the resulted improved.

Figure 7.2: The only misclassified instances to be removed from majority subset and the misclassified, borderline, noisy and weak instances to be removed from minority subset (circled dots *both blue and red*).
Figure 7.3: After applying PROMOTE: The only highly prominent synthetic minority instances resampled (*high density of minority instances in the squared box*).

### 7.4 Evaluating the algorithm

One method of judging the performance of a classifier is to compare the accuracy of all classifications. This number is calculated by using equation (7.1),

\[
\text{Accuracy} = \frac{(TP) + (TN)}{\text{TotalSamples}} \quad (7.1)
\]

Where, TP is the total number of correct positive classifications and TN is the total number of correct negative classifications, or correct rejections.

Another common method of evaluating classifier performances is to look at the sensitivity of a classification model. Sensitivity is equivalent to the true positive
rate, and is calculated as the number of true positive classifications divided by all positive classifications as given in equation (7.2),

\[
Sensitivity = \frac{TP}{(TP) + (FN)}
\] (7.2)

Another common metric in biomedical literature is the specificity of a classification model. Specificity, also known as the correct rejection rate, is defined as the number of true negative classifications divided by all negative classifications as given in equation (7.3),

\[
Specificity = \frac{TN}{(TN) + (FP)}
\] (7.3)

As a classifier becomes more sensitive it will identify a greater proportion of true positive instances, however, the number of false negative classifications will consequently rise. Similarly, as a classification model becomes more specific, i.e. correctly rejecting greater proportion of true negative instances, then the number of false positive classifications will also rise.

7.5 Experimental Results

Table 7.2: The evaluation results of C4.5 and PROMOTE for AUC, Precision, F-measure, Recall, Specificity and Accuracy.
From the experimental results reported in Table 7.2, Figure 7.1 and Figure 7.2, it is clear that our new algorithm PRMOTE can perform very well, especially if the objective is known. If accuracy is the main quality we are looking for then AUC and accuracy is required. Our new algorithm has greatly increased accuracy; since the AUC of our new algorithm over C4.5 is increased from $0.751\pm0.070$ to $0.846\pm0.043$ and accuracy of our new algorithm over C4.5 is increased from $74.49\pm5.27$ to $82.10\pm3.76$. When a good model of existing collected data is sought, then high Precision and F-measure is the objective. Our algorithm PRMOTE has improved the precision over C4.5 from $0.797\pm0.045$ to $0.821\pm0.044$. In case of F-measure also our algorithm PRMOTE is the best performing algorithm as the value is increased from $0.806\pm0.044$ to $0.833\pm0.037$ over C4.5.

For achieving good generalization over C4.5 our new algorithm recall and specificity measures have improved from $0.821\pm0.073$ to $0.850\pm0.061$ and
0.603±0.111 to 0.788±0.063 respectively. The best performing algorithm to predict outcome easily is again our new algorithm PRMOTE.

7.6 Conclusion

Early detection of diabetes disease is a challenging problem in medical domain for avoiding high death rates due to diabetes. We solve this problem in a scalable way by using our new classifier PRMOTE which gives simplified measures with high accuracy. Experiments conducted using a diabetes dataset collected during the study suggests that our algorithm have attained better results by showing supremacy in all the departments to generate a feasible solution for the early diabetes disease detection. The case study conducted on the UCI dataset shows the applicability of our algorithm PRMOTE in medical domain.