APPLICATIONS OF FEATURE SELECTION METHODS TO MEDICAL DATA

6.1 OVERVIEW

Electronic data repositories, especially in medical domains, contain enormous amounts of data. These data include currently unknown and potentially interesting patterns and relations, which can be uncovered using knowledge discovery and data mining methods [110]. The methods were successfully applied in a number of medical domains, e.g. in the localization of a primary tumor, prognostics of recurrence of breast cancer, diagnosis of thyroid diseases, rheumatology and Pulmonary Tuberculosis. The real Pulmonary Tuberculosis dataset is considered for this study. Tuberculosis (TB) is an infection caused by slow-growing bacteria that grow best in areas of the body that have lots of blood and oxygen. It is found in lungs. This is called as Pulmonary Tuberculosis. Tiredness, loss of weight, fever, fatigue, weight loss, coughing up blood, fever and night sweats, cough producing phlegm, wheezing, excessive sweating, Chest pain, breathing difficulty, etc. may be the symptoms associated with this disease. Some of these symptoms may be essential to diagnose the disease and some may not. Similarly, the laboratory test results are also considered to diagnose the disease. This chapter examines and compares the proposed EFRIMBS, EFAIMBS and NFMIMBS algorithms, which are discussed in the earlier chapters, for the diagnosis of real Pulmonary Tuberculosis dataset.

The rest of this chapter is organized as follows: Section 6.2 describes Pulmonary Tuberculosis disease. Section 6.3 illustrates the experimental
analysis of the proposed methods to the real time medical application and section 6.3 presents the summary of the chapter.

6.2 ABOUT PULMONARY TUBERCULOSIS

Pulmonary Tuberculosis (TB) is a contagious bacterial infection that involves the lungs, but may spread to other organs.

**Figure 6.1** Images of Lungs Affected by Pulmonary Tuberculosis
Causes

Pulmonary Tuberculosis (TB) is caused by the bacteria Mycobacterium tuberculosis. TB can be spread by breathing in air droplets from a cough or sneeze of an infected person. This is called primary TB. Nowadays, most people will recover from primary TB infection without further evidence of the disease. The infection may stay asleep or inactive (dormant) for years. However, in some people it can reactivate. Most people who develop symptoms of a TB infection first became infected in the past. However, in some cases, the disease may become active within weeks after the primary infection.

The elderly people, infants and people with weakened immune systems, for example due to Acquired Immune Deficiency Syndrome (AIDS), chemotherapy, diabetes, or certain medications are at higher risk for active TB. A person who has frequent contact with people who have TB or have poor nutrition or live in crowded or unsanitary living conditions may have high risk of contracting TB.

The following factors may increase the rate of TB infection in a population.

- Increase in Human Immunodeficiency Virus (HIV) infections
- Increase in number of homeless people (poor environment and nutrition)
- The appearance of drug-resistant strains of TB

However, rates vary dramatically by area of residence and socioeconomic status.
Symptoms

The primary stage of TB usually does not cause symptoms. When symptoms of pulmonary TB occur, the examination may show:

- Clubbing of the fingers or toes (in people with advanced disease)
- Enlarged or tender lymph nodes in the neck or other areas
- Fluid around a lung (pleural effusion)
- Unusual breath sounds (crackles)

Tests may include Biopsy of the affected tissue (rare), Bronchoscopy, Chest CT scan, Chest x-ray, Interferon-gamma blood test such as the QFT-Gold test to test for TB infection, Sputum examination and cultures, Thoracentesis and Tuberculin skin test.

Treatment

The goal of treatment is to cure the infection with drugs that fight the TB bacteria. Treatment of active pulmonary TB will always involve a combination of many drugs (usually four drugs). All the drugs are continued until lab tests show which medicines work best. The most commonly used drugs include Isoniazid, Rifampin, Pyrazinamide and Ethambutol. Other drugs that may be used to treat TB include Amikacin, Ethionamide, Moxifloxacin, Para-aminosalicylic acid and Streptomycin.

The patient may need to take many different pills at different times of the day for 6 months or longer. It is very important to take the pills in the way the health care provider instructed. When people do not take their TB medications as recommended, the infection becomes much more difficult to treat.
The TB bacteria may become resistant to treatment, and sometimes, the drugs no longer help to treat the infection. When there is a concern that a patient may not take all the medication as directed, a health care provider may need to watch the person take the prescribed drugs. This is called directly observed therapy. In this case, drugs may be given 2 or 3 times per week, as prescribed by a doctor. The patient may need to be admitted to a hospital for 2 - 4 weeks to avoid spreading the disease to others until he is no longer contagious.

**Prevention**

TB is a preventable disease, even in those who have been exposed to an infected person. Skin testing called Purified Protein Derivative (PPD) for TB is used in high risk populations or in people who may have been exposed to TB, such as health care workers.

A positive skin test indicates TB exposure and an inactive infection. In such cases the family doctor may be consulted for preventive therapy. People who have been exposed to TB should be skin tested immediately and have a follow-up test at a later date, if the first test is negative.

Prompt treatment is extremely important in controlling the spread of TB from those who have active TB disease to those who have never been infected with TB.

Some countries with a high incidence of TB give people a BCG vaccination to prevent TB. However, the effectiveness of this vaccine is controversial and it is not routinely used in the United States. People who have had Bacille Calmette Guerin (BCG) may still be skin tested for TB. The test results (if positive) should be discussed with the doctor.
6.3 EXPERIMENTAL ANALYSIS AND DISCUSSION

The proposed algorithms are implemented using MATLAB version 7.2 for real time medical database. The real Pulmonary Tuberculosis database consists of information collected directly from 200 patients at TB Center, Coimbatore, Tamil Nadu, India. It is a well-known center for diagnosis and treatment of TB. The advantage of this dataset is that it includes a sufficient number of records of different categories of people affected by TB. The set of descriptors presents all the required information about patients. The record of every patient contains many different attributes and this has been reduced to 20 attributes after consulting the Physician. The details of the attributes are given as follows: They are Age, Sex, Rising of temperature in the evening, Vomiting, Headache, Cough with expectoration, Abdominal pain, Diarrhoea, Hemoptyis, Loss of weight, Loss of appetite, Hoarseness of voice, Nasal block, Tiredness, Anemia, White cell count, Liver Function test, Erythrocyte sedimentation rate, Sputum Acid-Fast Bacillus (AFB) smear, Pulmonary parenchyma findings on X-ray and the decision attribute Diagnosis (Positive, Negative, Suspect).

The experiment is conducted for real time Pulmonary Tuberculosis database using three different proposed feature selection algorithms called EFRIMBS, EFAIMBS and NFMIMBS. These feature selection algorithms are applied to the dataset and consider only boundary samples to get minimum feature subset. The analysis is given below. First, EFRIMBS algorithm is applied to Pulmonary Tuberculosis database to select the feature subset.

The proposed EFRIMBS method is implemented for medical dataset Pulmonary Tuberculosis. This feature selection method selects minimum feature subset at minimum processing time and improves the classification
performance of the attributes selected by the proposed algorithm. Seven different classifiers LMT, Naive Bayes, SMO, C4.5, JRIP, PART and CART discussed in section 2.8 are used in this process.

In the proposed EFRIMBS method, K-Means discretization algorithm selects minimum number of features for feature subset at minimum processing time when compared to FCM and median as initial centroid of K-Means algorithms. The results of are discussed in Table 6.1. K-Means discretization gives higher classification accuracy rate to Naïve Bayes, PART and CART classifiers when compared to other classifiers. FCM based feature selection algorithm is also good to select features giving higher accuracy rate for C4.5, JRIP, PART and CART classifiers. Median as initial centroid of K-Means algorithm does not give minimum features for feature subset but it improves classification accuracy rate. The performance of the proposed EFRIMBS method for Pulmonary Tuberculosis with different classifiers is given in Figure 6.2.

Table 6.1 Results of EFRIMBS Method for Pulmonary Tuberculosis with Different Classifiers

<table>
<thead>
<tr>
<th>Discretization methods for Pulmonary Tuberculosis Datasets</th>
<th>Selected Number of features</th>
<th>Time (Seconds)</th>
<th>Classification Accuracy Rate of Classifiers in (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Raw Data</td>
<td></td>
<td>LMT</td>
</tr>
<tr>
<td>K-Means</td>
<td>20</td>
<td>06</td>
<td>10.30</td>
</tr>
<tr>
<td>FCM</td>
<td>20</td>
<td>07</td>
<td>12.90</td>
</tr>
<tr>
<td>Median as initial centroid of K-Means</td>
<td>20</td>
<td>19</td>
<td>20.59</td>
</tr>
</tbody>
</table>
Secondly, EFAIMBS algorithm is implemented on Pulmonary Tuberculosis Medical dataset. The proposed EFAIMBS algorithm gives reduced minimum feature subset for FCM discretization. The results are analysed in Table 6.2. With the selected minimum features at minimum processing time, it gives improved classification accuracy rate for FCM classification.

The classifiers JRIP, PART and CART give improved performance. Next to that K-Means algorithm is better to give minimum features for improved classification results of Naive Bayes, JRIP, PART and CART. But median as initial centroid of K-Means does not give minimum features for feature subset. The Performance of proposed EFAIMBS method for Pulmonary Tuberculosis with different classifiers is given in Figure 6.3.

Figure 6.2 Performance of the EFRIMBS Method for Pulmonary Tuberculosis with Different Classifiers
Table 6.2 Results of EFAIMBS Method for Pulmonary Tuberculosis with Different Classifiers

<table>
<thead>
<tr>
<th>Discretization method for Pulmonary Tuberculosis Datasets</th>
<th>Raw Data</th>
<th>Selected features</th>
<th>Time (Seconds)</th>
<th>Classification Accuracy Rate of Classifiers in (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LMT</td>
</tr>
<tr>
<td>K-Means</td>
<td>20</td>
<td>06</td>
<td>10.8</td>
<td>84.6</td>
</tr>
<tr>
<td>FCM</td>
<td>20</td>
<td>03</td>
<td>04.5</td>
<td>84.8</td>
</tr>
<tr>
<td>Median as initial centroid of K-Means</td>
<td>20</td>
<td>17</td>
<td>64.2</td>
<td>81.5</td>
</tr>
</tbody>
</table>

Figure 6.3 Performance of EFAIMBS Method for Pulmonary Tuberculosis with Different Classifiers

The performance of proposed NFMIMBS method for Pulmonary Tuberculosis with different classifiers is given in Figure 6.4.
Table 6.3 Results of NFMIMBS Method for Pulmonary Tuberculosis with Different Classifiers

<table>
<thead>
<tr>
<th>Discretization method for Pulmonary Tuberculosis Datasets</th>
<th>Raw Data</th>
<th>Selected features</th>
<th>Time (Seconds)</th>
<th>Classification Accuracy Rate of Classifiers in (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>LMT</td>
</tr>
<tr>
<td>K-Means</td>
<td>20</td>
<td>3</td>
<td>03.5</td>
<td>89.2</td>
</tr>
<tr>
<td>FCM</td>
<td>20</td>
<td>2</td>
<td>08.8</td>
<td>89.2</td>
</tr>
<tr>
<td>Median as initial centroid of K-Means</td>
<td>20</td>
<td>6</td>
<td>85.0</td>
<td>89.8</td>
</tr>
</tbody>
</table>

Figure 6.4 Performance of NFMIMBS Method for Pulmonary Tuberculosis with Different Classifiers

Finally, the proposed NFMIMBS method is used for feature subset selection. It gives improved classification accuracy rate for all K-Means, FCM and Median as initial centroid of K-Means feature selection algorithms. In this process, FCM discretization gives very minimum number of features for feature subset. From this analysis it is proved that
NFMIMBS feature selection method is better for almost all discretization methods. The FCM discretization of NFMIMBS gives improved performance to all classifiers. This proposed method also gives minimum features for median as initial centroid of K-Means with improved accuracy rate. In this process user defined threshold value $T_c = 0.1$ and $T_r = 0.5$ is used for all proposed methods to select features.

6.4 SUMMARY

The proposed feature selection algorithms are evaluated on the medical dataset Pulmonary Tuberculosis for different discretization algorithms. As medical dataset contains both numeric and nominal attributes, three different discretization algorithms namely K-Means, FCM and Median as initial centroid of K-Means are used to discretize the dataset. While comparing the performance of the proposed feature selection algorithms, EFRIMBS gives improved result for K-Means discretization. EFAIMBS gives better result for FCM discretization. The proposed NFMIMBS algorithm is good for all K-Means, FCM and Median as initial centroid of K-Means discretization.

To evaluate the subset of features seven different classifiers are used to improve its classification accuracy rate. By applying the proposed methods, the important features like Erythrocyte sedimentation rate, Sputum AFB smear and pulmonary parenchyma are selected for feature subset. Sometimes, in addition to the above selected features Cough with expectoration, Evening rise temperature and loss of appetite are also selected for feature subset. The results of the experiments are accepted by the domain experts as important feature for diagnosis of diseases. Thus the Feature selection method used here may support the diagnosis of diseases of Pulmonary Tuberculosis.