CHAPTER 3
DATA ACQUISITION

In this thesis, three new methods have been proposed for clustering. All these methods are tested with the data from medical domain, collected from the UCI repository. Totally 5 different medical data set has been used here. Some of the datasets contains missing attributes; those cases are eliminated from the dataset before subject to the experiments. This chapter presents a brief description about each data set and its attributes.

3.1 WISCONSIN BREAST CANCER DATABASE

The Wisconsin breast cancer database represents a reasonably complex problem with 9 continuous input attributes and two possible output classes. This data set was donated by William H. Wolberg of the University of Wisconsin hospitals. The dataset consists of 683 instances and we divided it into a training set of 341 instances and a test set of 342 instances each. The problem is to find whether the evidences indicate a benign or malignant neoplasm. The attributes for this data set are:

a) Radius (mean of distances from center to points on the perimeter)
b) Texture (standard deviation of gray-scale values)
c) Perimeter
d) Area
e) Smoothness (local variation in radius lengths)
f) Compactness (perimeter^2 / area - 1.0)
g) Concavity (severity of concave portions of the contour)
h) Concave points (number of concave portions of the contour)
i) symmetry
3.2 DERMATOLOGY DATABASE

This database contains 34 attributes, 33 of which are linear valued and one of them is nominal. The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope. In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. The names and id numbers of the patients were recently removed from the database. The attributes are:

Clinical Attributes: (take values 0, 1, 2, 3, unless otherwise indicated)

1. Erythema
2. Scaling
3. Definite borders
4. Itching
5. Koebner phenomenon
6. Polygonal papules
7. Follicular papules
8. Oral mucosal involvement
9. Knee and elbow involvement
10. Scalp involvement
11. Family history, (0 or 1)
12. Age (linear)

Histopathological Attributes: (take values 0, 1, 2, 3)
13. Melanin incontinence
14. Eosinophils in the infiltrate
15. PNL infiltrate
16. Fibrosis of the papillary dermis
17. Exocytosis
18. Acanthosis
19. Hyperkeratosis
20. Parakeratosis
21. Clubbing of the rete ridges
22. Elongation of the rete ridges
23. Thinning of the suprapapillary epidermis
24. Spongiform pustule
25. Unro microabcess
26. Focal hypergranulosis
27. Disappearance of the granular layer
28. Vacuolisation and damage of basal layer
29. Spongiosis
30. Saw-tooth appearance of retes
31. Follicular horn plug
32. Perifollicular parakeratosis
33. Inflammatory mononuclear infiltrate
34. Band-like infiltrate
3.3 HEPATITIS DATABASE

This dataset contains multivariate data with 155 instances on 19 attributes. The attributes of this data set are:

1. Class
2. Age
3. Sex
4. Steroid
5. Antivirals
6. Fatigue
7. Malaise
8. Anorexia
9. Liver Big
10. Liver Firm
11. Spleen Palpable
12. Spiders
13. Ascites
14. Varices
15. Bilirubin
16. Alk Phosphate
17. SGOT
18. Albumin
19. Prot ime
20. Histology

3.4 LUNG CANCER DATABASE

This data was used by Hong and Young to illustrate the power of the optimal discriminant plane even in ill-posed settings. Attribute 1 is the class label. All predictive attributes are nominal, taking on integer values 0-3. The data described 3
types of pathological lung cancers. This dataset contains 57 attributes with 32 instances.

3.5 THYROID DATABASE

The thyroid database was donated by Randolf Werner in 1992. It consists of 3772 learning examples and 3428 testing examples readily usable as ANN training and test sets. In the repository, these datasets are named: pub/machine-learning-databases/thyroid-disease/ann. Each example has 21 attributes, 15 of which are binary and 6 are continuous. The problem is to determine whether a patient referred to the clinic is hypothyroid. The output from the network is expected to be one of the three possible classes, namely: (i) normal (not hypothyroid), (ii) hyperfunction and (iii) subnormal function. In the dataset, 92 percent of the patients are not hyperthyroid and thus any reasonably good classifier should have above 92% correct predictions.

Data pre-processing involves preparing the sample data for further phases of the knowledge discovery from database process. This requires attention in removing the missing values and inconsistent values. If many records have missing values for key features then any data-mining method will not accurately discover patterns from those records. So we have eliminated the missing values and inconsistent values from the data set described above before getting it processed.