CHAPTER 2

Data Preprocessing

This chapter gives an overall introduction of the data preprocessing process, which cleans and optimizes data so that the data becomes ready for further data mining process.

2.1 DATA PREPROCESSING CONCEPTS

Data preprocessing describes the set of data processing procedures performed on raw data to prepare it for data mining. Raw data (sometimes called source data or atomic data) is the data that has not been processed for use. A distinction is sometimes made between data and information to the effect that information is the end product of data processing. Raw data has to undergo extraction, organization, etc., until it develops into cooked data to become useful information. Huge volume of raw data that is collected in a busy supermarket does not yield much information until it is processed. When it is processed, the data may indicate the particular items that each customer buys and the price at the time of purchase. Such information can be used for predictive analysis. It helps the owner to plan for future marketing campaigns. Data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. Data preprocessing is commonly used as a preliminary data mining practice, for example, in a neural network.

For preprocessing, there exist a number of different tools and methods which include sampling (from a large population of data, it selects a representative subset), transformation (produces a single input by manipulating raw data), de-noising (removes noise from data), normalization (organizes data for efficient access), and feature extraction (selects specified data that is significant in some particular context). In a customer relationship management context, data preprocessing

Scanned by CamScanner

is a component of Web mining. User transactions are generated by preprocessing of Web usage logs, from where meaningful sets of data can be extracted. User sessions may be identified by tracking the user, the requested websites and their order, and the length of time spent on each website. Once these have been pulled out of the raw data, they yield more useful information that can be put to the user's purposes, such as consumer research, marketing, or personalization. The raw data sometimes ends up in a database as a result of data processing. This data is accessible for further processing and analysis in a number of different ways.

Now consider the question "Why is data preprocessing required?" The real-world data that are to be analyzed by data mining techniques are as follows:

Incomplete: Incomplete data means missing of some attribute values or some interested attributes, or containing only aggregate data. Missing data, for tuples with missing values for some attributes, may need to be inferred.

Noisy: Data is said to be noisy if it contains errors or outlier values that deviate from the expected values. Noise refers to the modification of original values. For example, the distortion of a person's voice when talking on a poor-quality cell phone is a typical noise. Outliers are data points whose values are considerably different than majority of the other data objects in the set. In naming conventions or data code used, inconsistent formats for input fields, such as date inconsistencies, may result in making the incorrect data. It is necessary to implement some techniques to find and replace the noisy data.

Inconsistent: The data is said to be inconsistent if it contains discrepancies between different data items. For example, some attributes with same meaning may have different names in different databases. This will cause inconsistencies and redundancies. Naming inconsistencies may also occur for attribute values. For proper processing, we need to remove inconsistency in data.

Aggregate information: Aggregate information means something that is not the part of any precomputed data items in the data warehouse. If there is provision to obtain aggregate information such as sales per customer region, it will be more useful.

Enhancing mining process: The data mining process may become slow because of large number of data sets. So the performance of the mining process can be enhanced by reducing the number of data sets. It is also very important in data preprocessing.

Improve data quality: The quality of the data can be improved by data preprocessing techniques. These techniques subsequently improve the accuracy and efficiency of the mining process. Quality decisions must be based on quality data; so data preprocessing is an important step in the knowledge discovery process. Huge payoffs for decision making can be implemented by detecting data anomalies, rectifying them sooner, and reducing the data to be analyzed.

Now, we will discuss different forms of data processing. A pictorial representation

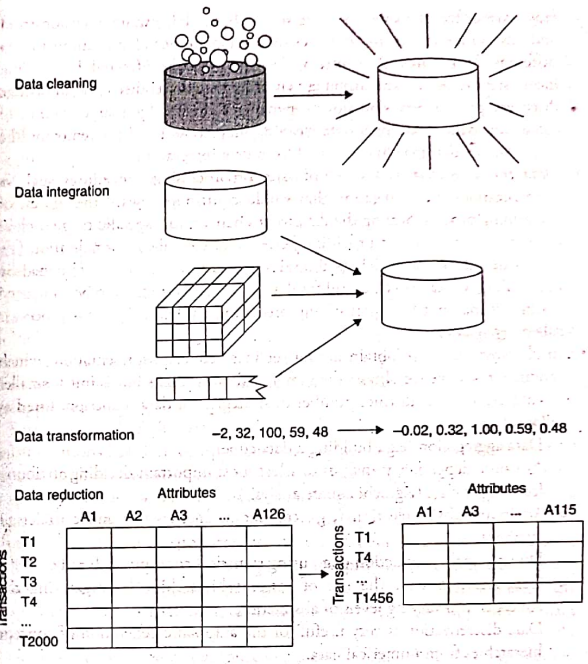

Fig. 2.1 Forms of data preprocessing

Sampling is an important phase of data selection in which we select a subset of the data set that has similar properties as the original data set. Statisticians sample because obtaining the entire set of data of interest and processing it is too expensive or time consuming.

Data cleaning is defined as the process of "cleaning" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Users are unlikely to trust the results of any data mining that has been applied to it if they believe the data are dirty. Dirty data may cause confusion for the data mining procedure or may result in unreliable output. But dirty data in a mining procedure are not always robust.

Data integration is defined as the process of integrating data from multiple databases, data cubes, or files. In a data mining environment, inconsistencies and redundancies are caused by attributes representing a given concept that

Scanned by CamScanner

may have different names in different databases. For example, customer_id and cust_id are the attribute names used for customer identification in two different data stores. Attribute values may also be affected by naming inconsistencies. Also, some attributes may be inferred from others. The knowledge discovery process may slow down or may get affected by a large amount of redundant data. Additional data cleaning may be required for removing the redundancies that may have resulted from data integration.

- *Data transformation* is the set of data preprocessing procedures such as normalization and aggregation that would contribute toward the success of the mining process. Scaling the data to be analyzed to a specific range such as a binary range [0, 1] for providing better results is called normalization [1]. Aggregate information such as the sales per customer region is obtained by data scaling, which will be useful for data analysis. It is needed to be computed because it may not be a part of any precomputed data cube. This process is called *aggregation*.
- In *data reduction*, we obtain a data set in a reduced representation, which produces the same (or almost the same) analytical results but is much smaller in volume. There is distinct number of strategies for data reduction, listed as follows:
  - Data aggregation (e.g., building a data cube)
  - Attribute dependency analysis to select most important deciding attributes for a target class (e.g., chi-square analysis)
  - Attribute subset selection (e.g., correlation analysis to remove irrelevant attributes)
  - Dimensionality reduction (e.g., using statistical concepts such as PCA)
  - Generalization with the use of concept hierarchies, by organizing the concepts into varying levels of abstraction
  - Data discretization is very useful for the automatic generation of concept hierarchies from numerical data.
- *Aggregation* is the process of combining two or more attributes (or objects) into a single attribute (or object). The main usages behind aggregation are data reduction (to reduce the number of attributes or objects), change of scale (cities aggregated into regions, states, countries, etc.), and to get more "stable" data as aggregated data tends to have less variability.

2.2 DATA CLEANING

Data cleaning is the technique used to improve the quality of the data by detecting and removing errors and inconsistencies from it. Data cleaning is also called data cleansing or scrubbing. Due to misspellings during data entry, missing information, or other invalid data, single data collections such as files and databases are affected by data quality problems.

The need for data cleaning increases significantly in data warehouses and other types of database systems where there is integration of multiple data sources.

To provide access to consistent, accurate data, the consolidation of different data representations and elimination of duplicate data becomes a mandate.

The correctness of the data in data warehouses is vital to avoid wrong conclusions, because data warehouses are used for decision making. Extensive support for data cleaning is provided by data warehouses.

The probability of the data being dirty is high in data warehouses because in the process called ETL (extract, transform, and load), they load and continuously refresh huge amounts of data from a variety of sources such as legacy data, files, and databases. Incorrect or misleading statistics ("garbage in, garbage out") is generated by duplicated or missing information [2].

Data cleaning is considered to be one of the hardest problems in data warehousing because of the wide range of possible data inconsistencies and the absolute data volume [1, 3]. Several requirements should be satisfied by the data cleaning approach. Both in individual data sources and when integrating multiple sources, all major errors and inconsistencies should be detected and removed by data cleaning. Tools to limit manual inspection and programming effort and to be extensible to easily cover additional sources should support this approach.

Furthermore, data cleaning is performed together with schema-related data transformations based on comprehensive metadata. It should not be performed in isolation. Data transformations and data cleaning mapping functions should be specified in such a way that other data sources as well as query processing phases may reuse these data transformations. In a reliable way, a workflow infrastructure should be supported to execute all data transformation steps for multiple sources for data warehouses. All the problems in data cleaning are closely related and, thus, should be treated in a uniform way. Any changes in the data structure, representation of data, or content of data are supported by data transformations.

Noise is a random error or variance in a measured variable. Following are some ways to handle noisy data:

Binning: Binning methods consult the "neighborhood," that is, the values around it, to smooth a sorted data value.

To distribute the sorted values, a number of buckets or bins are used. The binning method is said to perform local smoothing because binning methods consult the neighborhood of values.

In smoothing by bin means, the mean value of the bin is use to replace each value in a bin. Smoothing by bin medians can be implemented by using bin median to replace each bin value. In smoothing by bin boundaries, the bin boundaries are identified by the minimum and maximum values in a given bin. The closest boundary value is used to replace each bin value. In general, the greater the effect of smoothing, the larger is the width.

Alternatively, bins may be equal-width, where the interval range of values in each bin is constant. Binning is also used as a discretization technique. Binning is shown in Fig. 2.2

Scanned by CamScanner

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency)bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Fig. 2.2 Binning methods

Data cleaning is a process consisting of many types of operations, which as described as follows:

Discrepancy detection: Data entry errors, poorly designed data entry forms that have many optional fields, deliberate errors, data decay, and inconsistency in data representations are the main factors for data discrepancies. Any known knowledge of the properties of the data—called metadata—can be used for detecting discrepancy. Domain, data type, and acceptable values of each attribute are examples of metadata. The data trends are grasped and anomalies are identified by descriptive data summaries. When developers squeeze new attribute definitions into unused (bit) portions of already defined attributes (e.g., using an unused bit of an attribute whose value range uses only, say, 31 out of 32 bits), it causes another source of error, called field overloading. Unique rules, consecutive rules, and null rules are used to examine the data. Simple domain knowledge (e.g., knowledge of postal addresses and spell-checking) is used to detect errors and make corrections in the data by data scrubbing tools. When trying to clean the data obtained from multiple sources, data scrubbing tools rely on parsing and fuzzy matching techniques. Data auditing tools find discrepancies by analyzing the data to discover rules and relationships and detecting data that violate such conditions. They are the different data mining tool variants. For example, they may employ statistical analysis to find correlations or clustering to identify outliers. External references are used to find out some data inconsistencies that may be corrected manually.

Data transformation: Data transformation is required by most data mining applications, especially neural network based applications. It is the second step in data cleaning process. A series of data transformations is needed to be defined and applied to the data to correct the discrepancies, if found. Transformations are usually done on a batch basis. The user has to continuously check whether any discrepancies have been created by mistake after the completion of each transformation. Thus, the lack of interactivity is the greatest problem suffered by the entire data cleaning process. Increased interactivity is emphasized by new approaches to data cleaning. Metadata updation to reflect this knowledge is also an important thing. Future versions of the same data store will get speeded up by data cleaning process.

In general, data cleaning involves several phases:

Data analysis: In addition to a manual inspection of data samples, a detailed data analysis is required to detect which kinds of inconsistencies are to be cleaned out from the data set. To gain insight about the data properties and detect anomalies in the quality of data, analysis programs may be used.

Definition of transformation workflow and mapping rules: A large number of data transformation and cleaning steps may have to be executed depending on the number of data sources, their degree of heterogeneity, and the "dirtiness" of the data. A schema translation is used to map sources to a common data model. During early data cleaning steps, a relational representation is used by data warehouses such that single-source instance problems are corrected and the data can be prepared for integration.

For data warehousing, the workflow should be specified by the control and data flow for these transformation and cleaning steps. Declarative query and mapping language will specify the schema-related data transformations as well as the cleaning steps. Automatic generation of transformation code is enabled from this. Invoking user-written cleaning code and special purpose tools during a data transformation workflow is possible. User feedback on data instances for which they have no built-in cleaning logic is requested by transformation steps.

Verification: The testing and evaluation of correctness and effectiveness of the transformation workflow and the transformation definitions is required. The analysis, design, and verification steps may be iterated multiple times, if needed, e.g., after applying some transformations, some errors only become apparent.

Transformation: This happens as a part of the ETL phase or during answering queries on multiple sources.

Backflow of cleaned data: After transformation, the dirty data should be replaced in the original sources by cleaned data in order to give legacy applications the improved data. It also avoids redoing the cleaning work for future data extractions. For data warehousing, the data staging area provides the cleaned data.

Scanned by CamScanner

Large amount of metadata, such as schemas, instance-level data characteristics, transformation mappings, and workflow definitions are obviously required by the transformation process. This metadata should be maintained in a DBMS-based (database management system based) repository to employ consistency, flexibility, and ease of reuse. The details of the transformations applied to the original objects should be well documented and stored in the data warehouse.

2.3 HANDLING MISSING DATA

A common problem in statistical analysis is the missing data. Missing data rates of less than 1% are generally considered trivial, those of 1%–5% manageable, those between 5% and 15% require sophisticated methods to handle, and those more than 15% may seriously impact any kind of interpretation. In the literature to treat missing data, several methods have been proposed. To treat missing values in supervised classification problems, we can use four methods. They are case deletion (CD) technique, mean imputation (MI), median imputation (MDI), and *k*-nearest neighbor (KNN) imputation. The criterion to compare them is the effect on the misclassification rate of two classifiers—the linear discriminant analysis (LDA) and the KNN classifier. The first is a parametric classifier and the second one is a nonparametric classifier. The following four methods are there in a supervised classification context:

CD: It is also known as complete case analysis. Many software packages have a module that implements this technique. This method consists of all instances (cases) with missing values for at least one feature discarded. Determining the extent of missing data on each instance and attribute and deleting the instances and/or attributes with high levels of missing data is a variation of this method. It is necessary to evaluate the relevance of the attribute to the analysis before deleting it. Unfortunately, high degrees of missing values are associated with relevant attributes. If it involves minimal loss of sample size (minimal missing data or a sufficiently large sample size), then CD is less hazardous and there is no structure or pattern to the missing data. CD has been shown to produce more biased estimates than alternative methods for situations where the sample size is insufficient or some structure exists in the missing data. If the data are missing completely at random, then CD can be applied.

MI: This is one of the most frequently used methods. In this, the mean of all known values of that attribute in the class where the instance with missing attribute belongs will be calculated and it will be replaced for the missing data feature. Let us try to define it formally. Let us consider that the value X_{ij} of the kth class, C_k, is missing. Then it will be replaced by $X_{ij} = \Sigma(X_{ij}/nk)$, where nk represents the number of non-missing values in the jth feature of the kth class. We considered that the overall mean does not take into account the sample size of the class where the instance with the missing values belongs to. Some of the drawbacks of mean imputation are overestimation of sample size, underestimation

other types of database systems where there is integration of multiple data sources.

of covariance, and negatively biased correlation. Data sets used for supervised classification purposes give good experimental results for mean imputation.

Mode imputation: Since mean is affected by the presence of outliers, mode may be the next choice. Mode can be assumed as the most repeated value in a data sequence. As an example, consider the following scenario. Suppose we want to replace the missing value of the sex attribute of a record in a women's college. Naturally it has to be female. This is the notion behind the mode of a data sequence. Another possibility can be replacement with median which is called median imputation (MDI). It is also a widely adopted technique for missing value substitution.

The correlation structure of the data also has to be considered. Replacing missing data may become useless because of the existence of other features with similar information (high correlation) or similar predicting power.

KNN imputation (KNNI): In this method, a given number of instances that are most similar to the instance of interest are used to impute the missing value of an instance. A distance function is used to determine the similarity of two instances. The algorithm is as follows:

1. Data set D is divided into two parts. Instances in which at least one of the features is missing is contained in Dm. Dc is the set of instances which will have complete feature information.
2. For each vector x in Dm:
   (a) Instance vector divided into observed and missing parts as $x = [x_o; x_m]$.
   (b) Distance between x_o and all the instance vectors from the set Dc is calculated. Use only those features in the instance vectors from the complete set Dc that are observed in the vector x.
   (c) Perform a majority voting estimate of the missing values for categorical attributes by using the K closest instances vectors (K nearest neighbors).
3. The mean value of the attribute in the K-nearest neighborhood is used to replace the missing-value attribute. The median can be used instead of the mean.

The advantages of KNN imputation are as follows:

1. Qualitative attributes (the most frequent value among the K nearest neighbors) and quantitative attributes (the mean among the K nearest neighbors) are predicted by the K nearest neighbor.
2. Creating a predictive model for each attribute with missing data is not required. Actually, an explicit model is not created by the K-nearest neighbor algorithm.
3. Records with multiple missing values can be treated.

Following are a few other imputation methods that are also in use but are less popular:

shown in Fig. 2.2.

Scanned by CamScanner

Hot deck imputation: In this method, a missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data.

In random hot deck, an observed value (the donor) of the randomly chosen attribute is used to replace the missing value (the recipient) of an attribute.

Note: There are also cold deck imputation methods that are similar to hot deck. In this case, the data source to choose the imputed value must be different from the current data source.

Imputation using a prediction model: In this model, the missing data is substituted by estimated values that are predicted by the model. The attributes other than the missing data are used as input for the predictive model and the attribute with missing data is used as the response attribute.

The disadvantages of this approach are as follows:

1. The model estimated values are usually better-behaved than the true values.
2. If there are no relationships between the attributes in the data set and the attribute with missing data, then the model will not be precise for estimating missing values.
3. We have to build a large number of models to predict the missing values depending on the computational cost. Decision tree algorithms are used by imputation. In-built approaches are used to handle missing values in all the decision trees classifiers. The missing value of a given attribute is replaced by the corresponding value of a surrogate attribute that has the highest correlation with the original attribute. This approach is used in the CART (classification and regression tree) algorithm. A probabilistic approach is adopted in the C 4.5 algorithm to handle the missing data in both the training and the test sample.

Multiple imputations: In this method, for the treatment of missing values and their effect in the classifier accuracy feature, values drawn randomly (with replacement) from a fitted distribution are used to fill the missing values in a feature. This same step is repeated several times for accuracy, say $M = 15$ times. After that, the misclassification error for each data set is computed, by applying the classifier to each complete data set. A single estimation is obtained by averaging the misclassification error rates. Variances of the error rate are also estimated.

2.4 DATA TRANSFORMATION AND DISCRETIZATION

Some common types of transformations and normalizations will be discussed here.

In *attribute transformations*, for example, 0 may be used for male and 1 may be used for female. In decimal scaling, for the given data $X = \{2, 8, 10, 13, 17, 20\}$, applying the scaling transformation to the range [0, 1], {0.1, 0.4, 0.5, 0.65, 0.85, 1.0} is obtained by dividing with the maximum value 20. But the most popular

data normalization technique used is the *min–max normalization* where data is converted to a new range [0, 1] by applying common formula to all data [4]. It is sometimes called the change of origin and scale transformation where data values are changed by translating their origin by adding/subtracting a constant as well as by scaling down the range to another value. In *smoothing transformations*, the data values are adjusted to another value like in ceil(2.5) = 3, i.e., a rounding value transformation. In *scrubbing transformations*, whenever a replacement is needed, text strings are found and replaced. For example, if due to a data entry error, all names "Williams" were entered as "Bill," the entries can be corrected with a find–replace function.

2.5 DATA VISUALIZATION

Information that has been abstracted in some schematic form, including attributes or variables for the units of information, is called *data visualization*. This will represent the data visually. The structure of data that is to be mined is analyzed by data visualization. So data visualization is a very important task in data mining. The similarities, dissimilarities, clusters, and dependencies that exist in data are visualized by using this technique.

Weka is one of the most popular free and open-source software that is used nowadays for academic and research activities in data mining. It was developed by a team in the University of Waikato, New Zealand. It was developed as a collection of Java classes that are reusable, modifiable, and distributable. Almost all data mining algorithms/classifiers are available as Java classes and it is possible to develop new algorithms by modifying that source code. The Weka Knowledge Explorer is an easy-to-use graphical user interface that harnesses the power of the WEKA software. The explorer tool of WEKA allows various important data mining tasks such as classification and clustering to be manipulated graphically. The visualization tool helps to get a visual interpretation of the results of these tasks.

Along with the software package Weka, there are many standard data sets available, such as IRIS, CAR, and DIABATES, which are maintained by the UCI (University of California, Irvine, USA). These are the standard data sets that are used as benchmark for testing majority of new data mining algorithms in data mining research. Figure 2.3 shows the visualization of the standard data set named IRIS that is available with the Weka package. Figure 2.4 shows the distribution of data with respect to classes. It also shows the class distribution of the data. The visualize pane of the Weka package is the exclusive pane in Weka for data visualization. It can be used for analyzing various similarities and dependencies that are inherent in the data.

Attribute dependency analysis selects those attributes that are decisive for a target class. Chi-square analysis is a popular tool for this purpose. Finally dimensionality reduction is a phase in data preprocessing, which reduces the number of attributes in a data set. Concepts such as principal component

Scanned by CamScanner

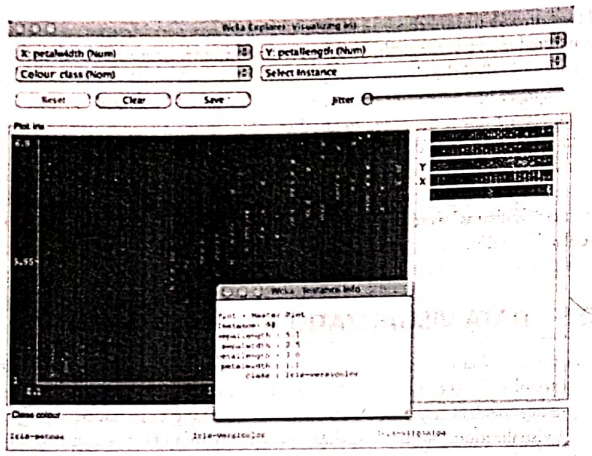

Fig. 2.3 Visualization of instance info

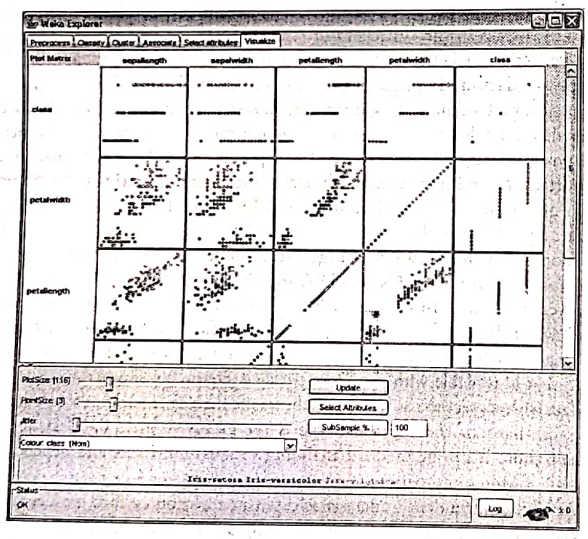

Fig. 2.4 Visualizing data in Weka software

other types of database systems where there is integration of multiple data

analysis and linear discriminant analysis are very popular dimensionality reduction techniques.

Summary statistics are numbers that summarize the various statistical properties of the data which give us insight into the nature of the data. Frequency, location (e.g., mean), and spread (e.g., standard deviation) are some examples. The percentage of times a value occurs in a data set is termed as its *frequency*. Consider the attribute "gender" and a sample of people. Suppose the gender "female" occurs about 50% of the time. Here female can be taken as the mode for the attribute gender. The most frequent attribute value is usually taken as mode. Categorical data uses the notions of frequency and mode.

SUMMARY

In this chapter, the basic concepts of data preprocessing were introduced and the importance of this process in data mining was understood. Before the actual process of data mining starts, the nature and structure of data is very important as it is to be checked against missing values, redundant data, attribute related problems, attribute dependencies, etc. If problems are found, the data has to be cleaned such that it can be submitted for data mining process.

Once the data is clean, it has to be submitted for transformation if needed. Transformation is all about transforming data to suit a particular data mining application. For example, neural networks work well when data is in 0/1 form. Hence, other numeric data ranges have to be transformed to this range. Data discretization using a technique called binning was also discussed, in which attribute values are smoothened using a frequency-based binning technique.

Finally, a software package WEKA has been introduced as one of the most popular data mining packages for academic and research purposes. It is a collection of reusable Java classes that can be recoded to develop new data mining classifiers. This software has built-in visualizing features that can be used to see the class distribution, similarities, clusters, etc., inherent in the data.

Data reduction techniques obtain a reduced representation of the data while minimizing the loss of information content. This may involve dimensionality reduction, which reduces the number of random variables or attributes under consideration. When the numbers of attributes are reduced, the data mining problem becomes much simpler to manage. Weka software packages help in many data preprocessing steps such as data visualization and attribute dependency analysis. Methods include wavelet transforms, principal component analysis, and attribute subset selection.

Different methods of data preprocessing have already been developed. But due to the huge amount of inconsistent or dirty data and the complexity of the problem, data preprocessing remains an active area of research.

Scanned by CamScanner

EXERCISES

Multiple Choice Questions

1. Which of the following is not a data cleaning technique?
   a. Binning b. Outlier detection
   c. Missing value replacement d. None of the above
2. The process of conversion of range of data is
   a. Data transformation b. Data cleaning
   c. Binning d. None of the above
3. Processing of data prior to data mining is called
   a. Data preprocessing b. Data collection
   c. Data reduction d. None of the above
4. MDI stands for
   a. Mean imputation b. Median imputation
   c. Maximal imputation d. None of the above
5. Reducing the number of attributes of data is
   a. Data analysis b. Data reduction
   c. Data preprocessing d. None of the above

Fill in the Blanks

1. _____ can be an example for a data discretizing technique.
2. One of the popular data transformation techniques is called _____.
3. _____ is a popular data preprocessing tool.
4. One of the powerful features of preprocessing in WEKA is _____.
5. One of the popular dimensionality reduction techniques is _____.

Explanatory Questions

1. Use min–max normalization to transform the value 35 for age to the range 0–1.
2. Explain the various data cleaning techniques.
3. Suppose a group of 12 marks has been sorted as follows:
   5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215
   Partition them into three bins by equal frequency partitioning.
4. Explain the various missing value handling techniques.
5. How can real world data be made ready for submitting to a neural network?

6. What is data visualization?
7. Explain the application of Weka in data preprocessing.
8. What is dimensionality reduction?
9. What is data integration?
10. What are the different methods that can be used for data smoothing?
11. Using the free software Weka, conduct a data visualization of the standard data sets such as IRIS and CAR, which are available with the software.
12. Can you suggest some popular software that can be used for preprocessing of data?
13. Write a brief note on the structure of the Weka software.

REFERENCES

1. J. Han, M. Kamber, J. Pei, *Data Mining Concepts and Techniques*, 3rd edition, MK publishers, 2012.
2. G. K. Guptha, *Introduction to Data Mining with Case Studies*, PHI publishers, 2006.
3. P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Pearson, 2006.
4. S. Elayidom, S. M. Idicula, and J. Alexander, *Comparison of Data Mining Techniques Based on Decision Trees and Neural Networks for Placement Chance Prediction*, Proceedings of International Conference ICONCEPT, Kerala, India, 2010.

ANSWER KEYS

Multiple Choice Questions

1. d 2. a 3. a 4. b 5. b

Fill in the Blanks

1. binning
2. normalization
3. SPSS
4. data visualization
5. principal component analysis

Scanned by CamScanner

Project Example

In a related research work, the training of neural network is done by data transformation. The inputs to the problem are particular rank, sex, reservation, sector, and branch of a student. The output is to predict the placement chances for that student. Inputs are fixed as categorical data and output has to be one from the set {Excellent, Good, Average, Poor}.

Usually, transformation operations involve converting categorical data to numeric data.

Normalization involves distributing the data evenly and scaling down into an acceptable range for the model. The details of the needed transformations are shown in Table 1.

Table 1 Attribute values mapped to 0 to 1 scale

Attribute	Range	Mapped to
RANK	1 to 4000	0 to 1
SEX	1 to 2	0 to 1
CATEGORY	1 to 4	0 to 1
SECTOR	1 to 2	0 to 1
BRANCH	A to J	0 to 1
ACTIVITY	1 to 4	One of the four values: E, G, A, and P.

For data normalization, (1) was used for transforming each data value D to I.

$$I = I_{min} + (I_{max} - I_{min}) \times \frac{(D - D_{min})}{(D_{max} - D_{min})} \quad (1)$$

Linear scaling requires the minimum and maximum values associated with each input. These values are D_{min} and D_{max}, respectively. $I_{min} - I_{max}$ is the input range required for the network. D_{min} and D_{max} are determined by the attribute values.

Since the neural network accepts input in the range either −1 to 1 or 0 to 1, all the input and output data are mapped to the data between 0 and 1. For example, the value 500 in the range {1–5000} will be transformed as 0 + (1 − 0) × (500 − 1)/(5000 − 1) = 0.099 in the range 0 to 1. Table 2 shows a snippet of sample data used to train neural network. Neural networks techniques are used by many data mining applications. All of them are provided categorical or numeric data as inputs; they require data transformations. Many built-in data preprocessing utilities are provided with many sophisticated data mining packages such as Weka.

The need for data cleaning [illegible] other types of database systems where there is integration of multiple data sources.

One of the most popular tools is SPSS (statistical package for social science problems) whose latest version is IBM SPSS Statistics 21.0 as on August, 2012. There are many add-ons such as SPSS missing values and SPSS data preparation, which are exclusive for data preprocessing. Another alternative is R Commander, which is an open-source alternative to SPSS based on the R programming language.

Table 2 Snippet of the sample data used to train the neural network

Sex	Reservation	Location	Rank	Branch
0	0	1	0.72	0.47
1	0	1	0.72	0.47
0	1	0	0.59	0.33
0	0	1	0.4	0.66
0	1	0	0.72	0.47
1	1	1	0.27	0.38

[illegible] shown in Fig. 2.2.

Scanned by CamScanner

CHAPTER 3

UCI Data Sets and Their Significance

This chapter gives an overall introduction of data sets from the UCI (University of California, Irvine) that are used as benchmarking datasets for testing new algorithms.

3.1 UCI DATA SETS

In data mining research for testing new algorithms, UCI (University of California, Irvine) data sets are the standard benchmark data sets used commonly. It is required that scientists prove the performance of the proposed algorithms over these standard data sets rather than over random data sets. They should show that their algorithms show better performances other than the standard algorithms in terms of accuracy, speed, or other statistics used for performance comparison. Usually, data mining research proceeds in different domains and researchers claim varying results across different domains. Hence, in order to standardize performance evaluation, these UCI data sets are widely used [1].

These UCI data sets are a result of efforts from data mining researchers for collecting data sets from various sources for their research purposes over many years. In this chapter, some important data sets used in machine learning are presented. The data in the data sets are partial in nature. In the case of some data sets, the number of data items is huge (of the order of thousands). But full versions are downloadable from [2].

3.1.1 Iris Plants Database

Source: Creator: R. A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, Annual Eugenics, 7, Part II, 179–188 (1936); also in *Contributions to*

Scanned by CamScanner

Relevant Information: This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day [3]. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two; the latter are not linearly separable from each other.

Predicted Attribute: Class of iris plant

Number of Instances: 150 (50 in each of three classes)

Number of Attributes: 4 numeric, predictive attributes, and the class

Attribute information:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm
5. Class:
   (a) *Iris setosa*
   (b) *Iris versicolour*
   (c) *Iris virginica*

Missing attribute values: None

Summary statistics:

	Min	Max	Mean	SD	Class correlation
Sepal length	4.3	7.9	5.84	0.83	0.7826
Sepal width	2.0	4.4	3.05	0.43	−0.4194
Petal length	1.0	6.9	3.76	1.76	0.9490 (high!)
Petal width	0.1	2.5	1.20	0.76	0.9565 (high!)

Class distribution: 33.3 for each of 3 classes.

```
@RELATION iris
@ATTRIBUTE sepallength   REAL
@ATTRIBUTE sepalwidth    REAL
@ATTRIBUTE petallength   REAL
@ATTRIBUTE petalwidth    REAL
@ATTRIBUTE class         {Iris-setosa, Iris-versicolor, Iris-virginica}
@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
```

other types of database systems where there is integration of multiple data sources.


```
4.7,3.2,1.3,0.2,Iris-setosa
4.6,3.1,1.5,0.2,Iris-setosa
5.0,3.6,1.4,0.2,Iris-setosa
5.4,3.9,1.7,0.4,Iris-setosa
4.6,3.4,1.4,0.3,Iris-setosa
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
5.1,3.5,1.4,0.3,Iris-setosa
5.7,3.8,1.7,0.3,Iris-setosa
5.1,3.8,1.5,0.3,Iris-setosa
5.4,3.4,1.7,0.2,Iris-setosa
5.1,3.7,1.5,0.4,Iris-setosa
4.6,3.6,1.0,0.2,Iris-setosa
5.1,3.3,1.7,0.5,Iris-setosa
4.8,3.4,1.9,0.2,Iris-setosa
5.0,3.0,1.6,0.2,Iris-setosa
5.0,3.4,1.6,0.4,Iris-setosa
5.2,3.5,1.5,0.2,Iris-setosa
5.2,3.4,1.4,0.2,Iris-setosa
4.7,3.2,1.6,0.2,Iris-setosa
4.8,3.1,1.6,0.2,Iris-setosa
5.4,3.4,1.5,0.4,Iris-setosa
5.2,4.1,1.5,0.1,Iris-setosa
5.5,4.2,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.0,3.2,1.2,0.2,Iris-setosa
5.5,3.5,1.3,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
4.4,3.0,1.3,0.2,Iris-setosa
5.1,3.4,1.5,0.2,Iris-setosa
5.0,3.5,1.3,0.3,Iris-setosa
4.5,2.3,1.3,0.3,Iris-setosa
4.4,3.2,1.3,0.2,Iris-setosa
5.0,3.5,1.6,0.6,Iris-setosa
5.1,3.8,1.9,0.4,Iris-setosa
4.8,3.0,1.4,0.3,Iris-setosa
```

shown in Fig. 2.2.

Scanned by CamScanner

5.1,3.8,1.6,0.2,Iris-setosa
4.6,3.2,1.4,0.2,Iris-setosa
5.3,3.7,1.5,0.2,Iris-setosa
5.0,3.3,1.4,0.2,Iris-setosa
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
5.0,2.0,3.5,1.0,Iris-versicolor
5.9,3.0,4.2,1.5,Iris-versicolor
6.0,2.2,4.0,1.0,Iris-versicolor
6.1,2.9,4.7,1.4,Iris-versicolor
5.6,2.9,3.6,1.3,Iris-versicolor
6.7,3.1,4.4,1.4,Iris-versicolor
5.6,3.0,4.5,1.5,Iris-versicolor
5.8,2.7,4.1,1.0,Iris-versicolor
6.2,2.2,4.5,1.5,Iris-versicolor
5.6,2.5,3.9,1.1,Iris-versicolor
5.9,3.2,4.8,1.8,Iris-versicolor
6.1,2.8,4.0,1.3,Iris-versicolor
6.3,2.5,4.9,1.5,Iris-versicolor
6.1,2.8,4.7,1.2,Iris-versicolor
6.4,2.9,4.3,1.3,Iris-versicolor
6.6,3.0,4.4,1.4,Iris-versicolor
6.8,2.8,4.8,1.4,Iris-versicolor
6.7,3.0,5.0,1.7,Iris-versicolor
6.0,2.9,4.5,1.5,Iris-versicolor
5.7,2.6,3.5,1.0,Iris-versicolor
5.5,2.4,3.8,1.1,Iris-versicolor
5.5,2.4,3.7,1.0,Iris-versicolor
5.8,2.7,3.9,1.2,Iris-versicolor
6.0,2.7,5.1,1.6,Iris-versicolor
5.4,3.0,4.5,1.5,Iris-versicolor
6.0,3.4,4.5,1.6,Iris-versicolor
6.7,3.1,4.7,1.5,Iris-versicolor
6.3,2.3,4.4,1.3,Iris-versicolor
5.6,3.0,4.1,1.3,Iris-versicolor
5.5,2.5,4.0,1.3,Iris-versicolor

5.5,2.6,4.4,1.2,Iris-versicolor
6.1,3.0,4.6,1.4,Iris-versicolor
5.8,2.6,4.0,1.2,Iris-versicolor
5.0,2.3,3.3,1.0,Iris-versicolor
5.6,2.7,4.2,1.3,Iris-versicolor
5.7,3.0,4.2,1.2,Iris-versicolor
5.7,2.9,4.2,1.3,Iris-versicolor
6.2,2.9,4.3,1.3,Iris-versicolor
5.1,2.5,3.0,1.1,Iris-versicolor
5.7,2.8,4.1,1.3,Iris-versicolor
6.3,3.3,6.0,2.5,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
7.1,3.0,5.9,2.1,Iris-virginica
6.3,2.9,5.6,1.8,Iris-virginica
6.5,3.0,5.8,2.2,Iris-virginica
7.6,3.0,6.6,2.1,Iris-virginica
4.9,2.5,4.5,1.7,Iris-virginica
7.3,2.9,6.3,1.8,Iris-virginica
6.7,2.5,5.8,1.8,Iris-virginica
7.2,3.6,6.1,2.5,Iris-virginica
6.5,3.2,5.1,2.0,Iris-virginica
6.4,2.7,5.3,1.9,Iris-virginica
6.8,3.0,5.5,2.1,Iris-virginica
5.7,2.5,5.0,2.0,Iris-virginica
5.8,2.8,5.1,2.4,Iris-virginica
6.4,3.2,5.3,2.3,Iris-virginica
6.5,3.0,5.5,1.8,Iris-virginica
7.7,3.8,6.7,2.2,Iris-virginica
7.7,2.6,6.9,2.3,Iris-virginica
6.0,2.2,5.0,1.5,Iris-virginica
6.9,3.2,5.7,2.3,Iris-virginica
5.6,2.8,4.9,2.0,Iris-virginica
7.7,2.8,6.7,2.0,Iris-virginica
6.3,2.7,4.9,1.8,Iris-virginica
6.7,3.3,5.7,2.1,Iris-virginica
7.2,3.2,6.0,1.8,Iris-virginica
6.2,2.8,4.8,1.8,Iris-virginica
6.1,3.0,4.9,1.8,Iris-virginica
6.4,2.8,5.6,2.1,Iris-virginica
7.2,3.0,5.8,1.6,Iris-virginica
7.4,2.8,6.1,1.9,Iris-virginica
7.9,3.8,6.4,2.0,Iris-virginica
6.4,2.8,5.6,2.2,Iris-virginica
6.3,2.8,5.1,1.5,Iris-virginica

Scanned by CamScanner

6.1,2.6,5.6,1.4,Iris-virginica
7.7,3.0,6.1,2.3,Iris-virginica
6.3,3.4,5.6,2.4,Iris-virginica
6.4,3.1,5.5,1.8,Iris-virginica
6.0,3.0,4.8,1.8,Iris-virginica
6.9,3.1,5.4,2.1,Iris-virginica
6.7,3.1,5.6,2.4,Iris-virginica
6.9,3.1,5.1,2.3,Iris-virginica
5.8,2.7,5.1,1.9,Iris-virginica
6.8,3.2,5.9,2.3,Iris-virginica
6.7,3.3,5.7,2.5,Iris-virginica
6.7,3.0,5.2,2.3,Iris-virginica
6.3,2.5,5.0,1.9,Iris-virginica
6.5,3.0,5.2,2.0,Iris-virginica
6.2,3.4,5.4,2.3,Iris-virginica
5.9,3.0,5.1,1.8,Iris-virginica

3.1.2 Diabetes Data Set

Sources: (a) Original owners: National Institute of Diabetes and Digestive and Kidney Diseases (b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)

Relevant information: Several constraints were placed on the selection of these instances from larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Number of instances: 768

Number of attributes: 8 plus class

For each attribute: (all numeric-valued)

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Missing attribute values: None

Class distribution: (Class value 1 is interpreted as "tested positive for diabetes")

other types of database systems [illegible]

Class value	Number of instances
0	500
1	268

Brief statistical analysis:

Attribute number	Mean	Standard deviation
1.	3.8	3.4
2.	120.9	32.0
3.	69.1	19.4
4.	20.5	16.0
5.	79.8	115.2
6.	32.0	7.9
7.	0.5	0.3
8.	33.2	11.8

Relabeled values in attribute 'class'
From: 0 To: tested_negative
From: 1 To: tested_positive

@relation pima_diabetes
@attribute 'preg' real
@attribute 'plas' real
@attribute 'pres' real
@attribute 'skin' real
@attribute 'insu' real
@attribute 'mass' real
@attribute 'pedi' real
@attribute 'age' real
@attribute 'class' {tested_negative, tested_positive}
note that a partial data set is shown below
@data
6,148,72,35,0,33.6,0.627,50,tested_positive
1,85,66,29,0,26.6,0.351,31,tested_negative
8,183,64,0,0,23.3,0.672,32,tested_positive
1,89,66,23,94,28.1,0.167,21,tested_negative
0,137,40,35,168,43.1,2.288,33,tested_positive
5,116,74,0,0,25.6,0.201,30,tested_negative
3,78,50,32,88,31,0.248,26,tested_positive
10,115,0,0,0,35.3,0.134,29,tested_negative
2,197,70,45,543,30.5,0.158,53,tested_positive

Scanned by CamScanner

8,125,96,0,0,0,0.232,54,tested_positive
4,110,92,0,0,37.6,0.191,30,tested_negative
10,168,74,0,0,38,0.537,34,tested_positive
10,139,80,0,0,27.1,1.441,57,tested_negative
1,189,60,23,846,30.1,0.398,59,tested_positive
5,166,72,19,175,25.8,0.587,51,tested_positive
7,100,0,0,0,30,0.484,32,tested_positive
0,118,84,47,230,45.8,0.551,31,tested_positive
7,107,74,0,0,29.6,0.254,31,tested_positive
1,103,30,38,83,43.3,0.183,33,tested_negative
1,115,70,30,96,34.6,0.529,32,tested_positive
3,126,88,41,235,39.3,0.704,27,tested_negative
8,99,84,0,0,35.4,0.388,50,tested_negative
7,196,90,0,0,39.8,0.451,41,tested_positive
9,119,80,35,0,29,0.263,29,tested_positive
11,143,94,33,146,36.6,0.254,51,tested_positive
10,125,70,26,115,31.1,0.205,41,tested_positive
7,147,76,0,0,39.4,0.257,43,tested_positive
1,97,66,15,140,23.2,0.487,22,tested_negative
13,145,82,19,110,22.2,0.245,57,tested_negative
5,117,92,0,0,34.1,0.337,38,tested_negative
5,109,75,26,0,36,0.546,60,tested_negative
3,158,76,36,245,31.6,0.851,28,tested_positive
3,88,58,11,54,24.8,0.267,22,tested_negative
6,92,92,0,0,19.9,0.188,28,tested_negative
10,122,78,31,0,27.6,0.512,45,tested_negative
4,103,60,33,192,24,0.966,33,tested_negative
11,138,76,0,0,33.2,0.42,35,tested_negative
9,102,76,37,0,32.9,0.665,46,tested_positive
2,90,68,42,0,38.2,0.503,27,tested_positive
4,111,72,47,207,37.1,1.39,56,tested_positive
3,180,64,25,70,34,0.271,26,tested_negative
7,133,84,0,0,40.2,0.696,37,tested_negative
7,106,92,18,0,22.7,0.235,48,tested_negative
9,171,110,24,240,45.4,0.721,54,tested_positive
7,159,64,0,0,27.4,0.294,40,tested_negative
0,180,66,39,0,42,1.893,25,tested_positive
1,146,56,0,0,29.7,0.564,29,tested_negative
2,71,70,27,0,28,0.586,22,tested_negative
7,103,66,32,0,39.1,0.344,31,tested_positive
7,105,0,0,0,0,0.305,24,tested_negative
1,103,80,11,82,19.4,0.491,22,tested_negative
1,101,50,15,36,24.2,0.526,26,tested_negative
5,88,66,21,23,24.4,0.342,30,tested_negative

8,176,90,34,300,33.7,0.467,58,tested_positive
7,150,66,42,342,34.7,0.718,42,tested_negative
1,73,50,10,0,23,0.248,21,tested_negative
7,187,68,39,304,37.7,0.254,41,tested_positive
0,100,88,60,110,46.8,0.962,31,tested_negative
0,146,82,0,0,40.5,1.781,44,tested_negative
0,105,64,41,142,41.5,0.173,22,tested_negative
2,84,0,0,0,0,0.304,21,tested_negative
8,133,72,0,0,32.9,0.27,39,tested_positive
5,44,62,0,0,25,0.587,36,tested_negative
2,141,58,34,128,25.4,0.699,24,tested_negative
7,114,66,0,0,32.8,0.258,42,tested_positive
5,99,74,27,0,29,0.203,32,tested_negative
0,109,88,30,0,32.5,0.855,38,tested_positive
2,109,92,0,0,42.7,0.845,54,tested_negative
1,95,66,13,38,19.6,0.334,25,tested_negative
4,146,85,27,100,28.9,0.189,27,tested_negative
2,100,66,20,90,32.9,0.867,28,tested_positive
5,139,64,35,140,28.6,0.411,26,tested_negative
13,126,90,0,0,43.4,0.583,42,tested_positive
4,129,86,20,270,35.1,0.231,23,tested_negative
1,79,75,30,0,32,0.396,22,tested_negative
1,0,48,20,0,24.7,0.14,22,tested_negative
7,62,78,0,0,32.6,0.391,41,tested_negative
5,95,72,33,0,37.7,0.37,27,tested_negative
0,131,0,0,0,43.2,0.27,26,tested_positive
2,112,66,22,0,25,0.307,24,tested_negative
3,113,44,13,0,22.4,0.14,22,tested_negative
2,74,0,0,0,0,0.102,22,tested_negative
7,83,78,26,71,29.3,0.767,36,tested_negative
0,101,65,28,0,24.6,0.237,22,tested_negative
5,137,108,0,0,48.8,0.227,37,tested_positive
2,110,74,29,125,32.4,0.698,27,tested_negative
13,106,72,54,0,36.6,0.178,45,tested_negative
2,100,68,25,71,38.5,0.324,26,tested_negative
15,136,70,32,110,37.1,0.153,43,tested_positive

3.1.3 Vehicle Data Set

Purpose

To classify a given silhouette as one of the four types of vehicles using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different angles.

Scanned by CamScanner

Problem Type

Classification

Source

Drs. Pete Mowforth and Barry Shepherd
Turing Institute
George House
36 North Hanover St.
Glasgow
G1 2AD

Contact

Alistair Sutherland
Statistics Dept.
Strathclyde University
Livingstone Tower
26 Richmond St.
GLASGOW G1 1XH
Great Britain
Tel: 041 552 4400 x3033
e-mail: alistair@uk.ac.strathclyde.stams

History

This data was originally gathered at the TI in 1986–87 by JP Siebert. It was partially financed by Barr and Stroud Ltd. The original purpose was to find a method of distinguishing three-dimensional objects within a two-dimensional image by the application of an ensemble of shape feature extractors to the two-dimensional silhouettes of the objects.

Description

The features were extracted from the silhouettes by the HIPS (hierarchical image processing system) extension BINATTS, which extracts a combination of scale-independent features utilizing both classical moments based measures, such as scaled variance, skewness, and kurtosis, about the major/minor axis and heuristic measures, such as hollows, circularity, rectangularity, and compactness. This particular combination of vehicles was chosen with the expectation that the bus, the van, and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars. The images were acquired by a camera looking downwards at the model vehicle from a fixed angle of elevation (34.2° to the horizontal). The vehicles were placed on a diffuse backlit surface (light box). They were painted matte black to minimize highlights. The images were captured using a CRS4000 frame store connected to a VAX 750. All images were captured with a spatial resolution of 128 × 128 pixels quantized to 64 gray levels.

other types of database systems where there is integration of multiple data sources.

Attributes

COMPACTNESS (average perim)**2/area
CIRCULARITY (average radius)**2/area
DISTANCE CIRCULARITY area/(av.distance from border)**2
RADIUS RATIO (max.rad-min.rad)/av.radius
PR.AXIS ASPECT RATIO (minor axis)/(major axis)
MAX.LENGTH ASPECT RATIO (length perp. max length)/(max length)
SCATTER RATIO (inertia about minor axis)/(inertia about major axis)
ELONGATEDNESS area/(shrink width)**2
PR.AXIS RECTANGULARITY area/(pr.axis length*pr.axis width)
MAX.LENGTH RECTANGULARITY area/(max.length*length perp. to this)
SCALED VARIANCE (2nd order moment about minor axis)/area ALONG MAJOR AXIS
SCALED VARIANCE (2nd order moment about major axis)/area ALONG MINOR AXIS
SCALED RADIUS OF GYRATION (mavar+mivar)/area
SKEWNESS ABOUT (3rd order moment about major axis)/sigma_min**3 MAJOR AXIS
SKEWNESS ABOUT (3rd order moment about minor axis)/sigma_maj**3 MINOR AXIS
KURTOSIS ABOUT (4th order moment about major axis)/sigma_min**4 MINOR AXIS
KURTOSIS ABOUT (4th order moment about minor axis)/sigma_maj**4 MAJOR AXIS
HOLLOWS RATIO (area of hollows)/(area of bounding polygon)

where sigma_maj**2 is the variance along the major axis and sigma_min**2 is the variance along the minor axis, and

area of hollows = area of bounding polygon – area of object

The area of the bounding polygon is found as a side result of the computation to find the maximum length. Each individual length computation yields a pair of calipers to the object orientated at every 5 degrees. The object is propagated into an image containing the union of these calipers to obtain an image of the bounding polygon.

Number of Classes: 4: opel, saab, bus, van

Scanned by CamScanner

Number of Examples
Total no. = 946
No. in each class:
opel—240
saab—240
bus—240
van—226

Number of Attributes:
No. of attributes. = 18
@relation vehicle
@attribute 'COMPACTNESS' real
@attribute 'CIRCULARITY' real
@attribute 'DISTANCE CIRCULARITY' real
@attribute 'RADIUS RATIO' real
@attribute 'PR.AXIS ASPECT RATIO' real
@attribute 'MAX.LENGTH ASPECT RATIO' real
@attribute 'SCATTER RATIO' real
@attribute 'ELONGATEDNESS' real
@attribute 'PR.AXIS RECTANGULARITY' real
@attribute 'MAX.LENGTH RECTANGULARITY' real
@attribute 'SCALED VARIANCE_MAJOR' real
@attribute 'SCALED VARIANCE_MINOR' real
@attribute 'SCALED RADIUS OF GYRATION' real
@attribute 'SKEWNESS ABOUT_MAJOR' real
@attribute 'SKEWNESS ABOUT_MINOR' real
@attribute 'KURTOSIS ABOUT_MAJOR' real
@attribute 'KURTOSIS ABOUT_MINOR' real
@attribute 'HOLLOWS RATIO' real
@attribute 'Class' {opel,saab,bus,van}
Note that a partial data set is shown below:
@data
95,48,83,178,72,10,162,42,20,159,176,379,184,70,6,16,187,197,van
91,41,84,141,57,9,149,45,19,143,170,330,158,72,9,14,189,199,van
104,50,106,209,66,10,207,32,23,158,223,635,220,73,14,9,188,196,saab
93,41,82,159,63,9,144,46,19,143,160,309,127,63,6,10,199,207,van
85,44,70,205,103,52,149,45,19,144,241,325,188,127,9,11,180,183,bus
107,57,106,172,50,6,255,26,28,169,280,957,264,85,5,9,181,183,bus
97,43,73,173,65,6,153,42,19,143,176,361,172,66,13,1,200,204,bus
90,43,66,157,65,9,137,48,18,146,162,281,164,67,3,3,193,202,van
86,34,62,140,61,7,122,54,17,127,141,223,112,64,2,14,200,208,van
93,44,98,197,62,11,183,36,22,146,202,505,152,64,4,14,195,204,saab
86,36,70,143,61,9,133,50,18,130,153,266,127,66,2,10,194,202,van
90,34,66,136,55,6,123,54,17,118,148,224,118,65,5,26,196,202,saab

88,46,74,171,68,6,152,43,19,148,180,349,192,71,5,11,189,195,bus
89,42,85,144,58,10,152,44,19,144,173,345,161,72,8,13,187,197,van
94,49,79,203,71,5,174,37,21,154,196,465,206,71,6,2,197,199,bus
96,55,103,201,65,9,204,32,23,166,227,624,246,74,6,2,186,194,opel
89,36,51,109,52,6,118,57,17,129,137,206,125,80,2,14,181,185,van
99,41,77,197,69,6,177,36,21,139,202,485,151,72,4,10,198,199,bus
104,54,100,186,61,10,216,31,24,173,225,686,220,74,5,11,185,195,saab
101,56,100,215,69,10,208,32,24,169,227,651,223,74,6,5,186,193,opel
84,47,75,153,64,6,154,43,19,145,175,354,184,75,0,3,185,192,bus
84,37,53,121,59,5,123,55,17,125,141,221,133,82,7,1,179,183,van
94,43,64,173,69,7,150,43,19,142,169,344,177,68,9,1,199,206,bus
87,39,70,148,61,7,143,46,18,136,164,307,141,69,1,2,192,199,bus
99,53,105,219,66,11,204,32,23,165,221,623,224,68,0,6,191,201,saab
85,45,80,154,64,9,147,45,19,148,169,324,174,71,1,4,188,199,van
83,36,54,119,57,6,128,53,18,125,143,238,139,82,6,3,179,183,saab
107,54,98,203,65,11,218,31,25,167,229,696,216,72,1,28,187,199,saab
102,45,85,193,64,6,192,33,22,146,217,570,163,76,6,7,195,193,bus
80,38,63,129,55,7,146,46,19,130,168,314,158,83,9,20,180,185,saab
89,43,85,160,64,11,155,43,19,151,173,356,174,72,5,9,185,196,van
88,42,77,151,58,8,140,47,18,142,165,293,158,64,10,11,198,205,saab
93,35,66,154,59,6,142,46,18,128,162,304,120,64,5,13,197,202,opel
101,48,107,222,68,10,208,32,24,154,232,641,204,70,5,38,190,202,opel
87,38,85,177,61,8,164,40,20,129,186,402,130,63,1,25,198,205,opel
100,46,90,172,67,9,157,43,20,150,170,363,184,67,17,7,192,200,van
82,44,72,118,52,7,152,44,19,147,174,340,177,82,2,2,180,185,bus
90,48,86,306,126,49,153,44,19,156,272,346,200,118,0,15,185,194,van
106,53,98,176,54,10,216,31,24,171,235,691,218,74,1,9,187,197,saab
81,45,68,169,73,6,151,44,19,146,173,336,186,75,7,0,183,189,bus
95,48,104,214,67,9,205,32,23,151,227,628,202,74,5,9,186,193,opel
88,37,51,105,52,5,119,57,17,128,135,207,125,86,8,16,179,183,van
94,49,87,137,54,11,158,43,20,162,178,366,186,75,5,5,183,194,van
93,37,76,183,63,8,164,40,20,134,191,405,139,57,4,7,192,197,saab
119,54,106,220,65,12,213,31,24,167,223,675,232,66,20,1,192,202,saab
93,46,82,145,58,11,159,43,20,160,180,371,189,77,2,4,183,194,van
91,43,70,133,55,8,130,51,18,146,159,253,156,70,1,8,190,194,van
85,42,66,122,54,6,148,46,19,141,172,317,174,88,6,14,180,182,bus
89,47,81,147,64,11,156,44,20,163,170,352,188,76,6,13,184,193,van
91,45,79,176,59,9,163,40,20,148,184,404,179,62,0,10,199,208,saab
78,38,63,115,51,6,142,47,19,130,162,299,146,77,2,4,181,185,saab
92,38,71,174,66,7,154,43,19,133,181,355,130,70,4,24,189,195,saab
98,55,101,228,70,9,210,31,24,168,236,661,245,72,1,6,188,197,opel
101,42,62,175,67,6,149,43,19,139,169,341,165,65,7,11,202,209,bus
101,56,104,185,53,6,257,26,28,168,275,956,230,83,5,26,180,184,bus
94,36,66,151,61,8,133,50,18,135,154,265,119,62,9,3,201,208,van

Scanned by CamScanner

97,44,96,195,63,9,185,36,22,144,202,512,165,66,4,8,191,199,saab
89,47,84,133,55,11,157,44,20,160,169,354,176,74,5,9,182,192,van
107,53,103,221,66,11,209,32,24,163,222,653,212,66,0,1,191,201,opel
85,39,68,119,52,5,128,53,18,135,148,241,142,75,8,8,182,187,van
103,50,98,212,63,9,193,34,22,161,214,567,185,64,5,5,198,204,opel
77,38,63,135,59,5,130,52,18,130,145,247,139,79,13,21,183,187,opel
96,40,70,120,50,8,137,50,18,141,162,269,139,80,10,13,183,183,van
83,42,66,156,67,7,150,45,19,144,174,333,159,78,4,2,182,188,bus
93,45,86,201,69,7,184,35,22,145,203,523,183,72,0,4,194,197,bus
89,41,75,143,56,7,146,46,19,137,170,317,156,76,18,5,184,188,opel
81,43,68,125,57,8,149,46,19,146,169,323,172,83,6,18,179,184,bus
98,55,101,219,69,11,225,30,25,178,231,748,216,74,6,14,187,195,opel
86,44,78,164,68,9,142,46,18,147,168,305,171,70,1,11,190,201,van
98,49,84,219,74,7,190,34,22,154,208,558,209,74,4,7,195,195,bus
96,55,98,161,54,10,215,31,24,175,226,683,221,76,3,6,185,193,opel
97,59,108,227,70,11,224,30,25,186,225,732,218,70,10,25,186,198,opel
92,39,91,191,62,8,176,37,21,137,196,466,151,67,3,23,192,200,opel
84,38,60,128,56,5,132,50,18,130,148,261,141,75,8,4,185,188,opel
89,44,80,191,66,6,162,40,20,143,189,396,180,66,13,11,194,199,opel
93,47,84,205,71,7,176,36,21,152,190,476,201,70,7,19,198,201,bus
104,52,101,206,62,10,198,33,23,161,207,587,204,64,2,5,195,204,opel
81,43,68,148,64,7,150,45,19,144,175,330,171,80,1,2,182,185,bus
88,45,82,155,56,8,154,43,19,149,180,357,170,69,3,0,188,193,saab
103,51,105,174,56,11,210,32,24,163,222,650,222,73,8,9,187,196,saab
83,46,68,139,59,6,150,44,19,146,172,336,183,74,5,3,185,191,bus
79,38,55,120,55,5,142,48,19,128,153,295,145,81,4,2,180,183,saab
97,55,103,197,63,11,215,31,24,172,219,677,219,75,5,24,185,194,opel
83,39,69,127,54,5,135,49,18,131,155,274,162,69,16,6,187,190,opel
98,38,72,192,69,5,166,38,20,131,189,427,138,70,1,3,200,202,bus
90,48,77,132,56,10,157,44,20,164,169,354,187,78,1,3,182,191,van
85,43,66,123,55,7,150,45,19,146,172,326,173,83,4,15,180,183,bus
81,45,68,154,69,22,151,45,19,147,186,335,186,88,1,10,180,185,bus
90,48,85,157,64,11,161,43,20,167,175,375,186,74,3,16,185,195,van
104,53,108,204,64,11,220,31,25,172,226,707,203,71,14,30,189,203,saab
95,43,96,202,65,10,189,35,22,143,217,534,166,71,6,27,190,197,opel
93,42,98,192,63,9,185,36,22,138,206,508,173,70,10,21,189,197,saab
87,38,71,123,53,8,137,49,18,127,158,277,145,75,0,9,181,186,saab
104,56,96,231,74,11,220,30,25,172,223,713,218,73,6,16,186,195,opel
95,41,82,170,65,9,145,46,19,145,163,314,140,64,4,8,199,207,van
105,54,105,213,67,10,200,33,23,163,214,597,214,68,10,20,190,198,opel
106,55,96,196,60,12,221,30,25,173,225,717,214,72,9,13,186,196,opel
86,39,84,149,57,8,156,43,20,133,185,358,157,74,0,23,183,190,opel
95,49,92,193,62,10,178,37,21,154,200,478,171,64,2,0,198,206,opel
99,57,100,177,54,13,224,30,25,188,223,726,213,72,4,7,185,198,opel
89,42,66,125,53,7,131,51,18,144,162,254,162,73,10,17,188,191,van

other types of database systems where these

95,49,82,139,56,11,159,43,20,162,173,365,185,75,7,10,182,191,van
97,37,70,173,66,7,151,43,19,129,167,346,119,65,0,16,201,208,bus
100,47,70,185,70,7,162,40,20,153,179,406,172,68,9,6,200,205,bus
108,49,109,204,61,11,212,31,24,159,229,665,215,71,16,11,190,199,saab
92,46,83,154,56,6,160,41,20,148,185,382,184,71,10,5,186,191,saab
82,36,51,114,53,4,135,50,18,126,150,268,144,86,15,4,181,182,saab
111,58,105,183,51,6,265,26,29,174,285,1018,255,85,4,8,181,183,bus
87,45,66,139,58,8,140,47,18,148,168,294,175,73,3,12,188,196,van
94,46,77,169,60,8,158,42,20,148,181,373,181,67,12,2,193,199,saab
95,43,76,142,57,10,151,44,19,149,173,339,159,71,2,23,187,200,van
90,44,72,157,64,8,137,48,18,144,159,283,171,65,9,4,196,203,van
93,34,66,140,56,7,130,51,18,120,151,251,114,62,5,29,201,207,opel
93,39,87,183,64,8,169,40,20,134,200,422,149,72,7,25,188,195,saab
89,46,84,163,66,11,159,43,20,159,173,368,176,72,1,20,186,197,van
106,54,101,222,67,12,222,30,25,173,228,721,200,70,3,4,187,201,saab
86,36,78,146,58,7,135,50,18,124,155,270,148,66,0,25,190,195,saab
85,36,66,123,55,5,120,56,17,128,140,212,131,73,1,18,186,190,van

3.1.4 Wine Data Set

Sources: (a) M. Forina, et al, *PARVUS—An Extendible Package for Data Exploration, Classification, and Correlation*, Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy. (b) Stefan Aeberhard, email: stefan@coral.cs.jcu.edu.au

Relevant information: These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

The attributes are

1. Alcohol
2. Malic acid
3. Ash
4. Alkalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

Scanned by CamScanner

Number of Instances:
Class 1—59
Class 2—71
Class 3—48

Number of Attributes: 13

For Each Attribute:
1. All attributes are continuous
2. No statistics available, but suggest to standardize variables for certain uses (e.g., for us with classifiers which are NOT scale invariant)

Note: The first attribute is class identifier (1–3)

Missing Attribute Values: None

Class Distribution: Number of instances per class
Class 1–59
Class 2–71
Class 3–48

Information about the Data Set:
CLASSTYPE: nominal
CLASSINDEX: first

@relation wine
@attribute class {1,2,3}
@attribute Alcohol REAL
@attribute Malic_acid REAL
@attribute Ash REAL
@attribute Alkalinity_of_ash REAL
@attribute Magnesium INTEGER
@attribute Total_phenols REAL
@attribute Flavanoids REAL
@attribute Nonflavanoid_phenols REAL
@attribute Proanthocyanins REAL
@attribute Color_intensity REAL
@attribute Hue REAL
@attribute OD280/OD315_of_diluted_wines REAL
@attribute Proline INTEGER

@data
1,14.23,1.71,2.43,15.6,127,2.8,3.06,.28,2.29,5.64,1.04,3.92,1065
1,13.2,1.78,2.14,11.2,100,2.65,2.76,.26,1.28,4.38,1.05,3.4,1050
1,13.16,2.36,2.67,18.6,101,2.8,3.24,.3,2.81,5.68,1.03,3.17,1185
1,14.37,1.95,2.5,16.8,113,3.85,3.49,.24,2.18,7.8,.86,3.45,1480
1,13.24,2.59,2.87,21,118,2.8,2.69,.39,1.82,4.32,1.04,2.93,735

1,14.2,1.76,2.45,15.2,112,3.27,3.39,.34,1.97,6.75,1.05,2.85,1450
1,14.39,1.87,2.45,14.6,96,2.5,2.52,.3,1.98,5.25,1.02,3.58,1290
1,14.06,2.15,2.61,17.6,121,2.6,2.51,.31,1.25,5.05,1.06,3.58,1295
1,14.83,1.64,2.17,14,97,2.8,2.98,.29,1.98,5.2,1.08,2.85,1045
1,13.86,1.35,2.27,16,98,2.98,3.15,.22,1.85,7.22,1.01,3.55,1045
1,14.1,2.16,2.3,18,105,2.95,3.32,.22,2.38,5.75,1.25,3.17,1510
1,14.12,1.48,2.32,16.8,95,2.2,2.43,.26,1.57,5,1.17,2.82,1280
1,13.75,1.73,2.41,16,89,2.6,2.76,.29,1.81,5.6,1.15,2.9,1320
1,14.75,1.73,2.39,11.4,91,3.1,3.69,.43,2.81,5.4,1.25,2.73,1150
1,14.38,1.87,2.38,12,102,3.3,3.64,.29,2.96,7.5,1.2,3,1547
1,13.63,1.81,2.7,17.2,112,2.85,2.91,.3,1.46,7.3,1.28,2.88,1310
1,14.3,1.92,2.72,20,120,2.8,3.14,.33,1.97,6.2,1.07,2.65,1280
1,13.83,1.57,2.62,20,115,2.95,3.4,.4,1.72,6.6,1.13,2.57,1130
1,14.19,1.59,2.48,16.5,108,3.3,3.93,.32,1.86,8.7,1.23,2.82,1680
1,13.64,3.1,2.56,15.2,116,2.7,3.03,.17,1.66,5.1,.96,3.36,845
1,14.06,1.63,2.28,16,126,3,3.17,.24,2.1,5.65,1.09,3.71,780
1,12.93,3.8,2.65,18.6,102,2.41,2.41,.25,1.98,4.5,1.03,3.52,770
1,13.71,1.86,2.36,16.6,101,2.61,2.88,.27,1.69,3.8,1.11,4,1035
1,12.85,1.6,2.52,17.8,95,2.48,2.37,.26,1.46,3.93,1.09,3.63,1015
1,13.5,1.81,2.61,20,96,2.53,2.61,.28,1.66,3.52,1.12,3.82,845
1,13.05,2.05,3.22,25,124,2.63,2.68,.47,1.92,3.58,1.13,3.2,830
1,13.39,1.77,2.62,16.1,93,2.85,2.94,.34,1.45,4.8,.92,3.22,1195
1,13.3,1.72,2.14,17,94,2.4,2.19,.27,1.35,3.95,1.02,2.77,1285
1,13.87,1.9,2.8,19.4,107,2.95,2.97,.37,1.76,4.5,1.25,3.4,915
1,14.02,1.68,2.21,16,96,2.65,2.33,.26,1.98,4.7,1.04,3.59,1035
1,13.73,1.5,2.7,22.5,101,3,3.25,.29,2.38,5.7,1.19,2.71,1285
1,13.58,1.66,2.36,19.1,106,2.86,3.19,.22,1.95,6.9,1.09,2.88,1515
1,13.68,1.83,2.36,17.2,104,2.42,2.69,.42,1.97,3.84,1.23,2.87,990
1,13.76,1.53,2.7,19.5,132,2.95,2.74,.5,1.35,5.4,1.25,3,1235
1,13.51,1.8,2.65,19,110,2.35,2.53,.29,1.54,4.2,1.1,2.87,1095
1,13.48,1.81,2.41,20.5,100,2.7,2.98,.26,1.86,5.1,1.04,3.47,920
1,13.28,1.64,2.84,15.5,110,2.6,2.68,.34,1.36,4.6,1.09,2.78,880
1,13.05,1.65,2.55,18,98,2.45,2.43,.29,1.44,4.25,1.12,2.51,1105
1,13.07,1.5,2.1,15.5,98,2.4,2.64,.28,1.37,3.7,1.18,2.69,1020
1,14.22,3.99,2.51,13.2,128,3,3.04,.2,2.08,5.1,.89,3.53,760
1,13.56,1.71,2.31,16.2,117,3.15,3.29,.34,2.34,6.13,.95,3.38,795
1,13.41,3.84,2.12,18.8,90,2.45,2.68,.27,1.48,4.28,.91,3,1035
1,13.88,1.89,2.59,15,101,3.25,3.56,.17,1.7,5.43,.88,3.56,1095
1,13.24,3.98,2.29,17.5,103,2.64,2.63,.32,1.66,4.36,.82,3,680
1,13.05,1.77,2.1,17,107,3,3,.28,2.03,5.04,.88,3.35,885
1,14.21,4.04,2.44,18.9,111,2.85,2.65,.3,1.25,5.24,.87,3.33,1080
1,14.38,3.59,2.28,16,102,3.25,3.17,.27,2.19,4.9,1.04,3.44,1065
1,13.9,1.68,2.12,16,101,3.1,3.39,.21,2.14,6.1,.91,3.33,985
1,14.1,2.02,2.4,18.8,103,2.75,2.92,.32,2.38,6.2,1.07,2.75,1060

Scanned by CamScanner

1,13.94,1.73,2.27,17.4,108,2.88,3.54,.32,2.08,8.90,1.12,3.1,1260
1,13.05,1.73,2.04,12.4,92,2.72,3.27,.17,2.91,7.2,1.12,2.91,1150
1,13.83,1.65,2.6,17.2,94,2.45,2.99,.22,2.29,5.6,1.24,3.37,1265
1,13.82,1.75,2.42,14,111,3.88,3.74,.32,1.87,7.05,1.01,3.26,1190
1,13.77,1.9,2.68,17.1,115,3,2.79,.39,1.68,6.3,1.13,2.93,1375
1,13.74,1.67,2.25,16.4,118,2.6,2.9,.21,1.62,5.85,.92,3.2,1060
1,13.56,1.73,2.46,20.5,116,2.96,2.78,.2,2.45,6.25,.98,3.03,1120
1,14.22,1.7,2.3,16.3,118,3.2,3,.26,2.03,6.38,.94,3.31,970
1,13.29,1.97,2.68,16.8,102,3,3.23,.31,1.66,6,1.07,2.84,1270
1,13.72,1.43,2.5,16.7,108,3.4,3.67,.19,2.04,6.8,.89,2.87,1285
2,12.37,.94,1.36,10.6,88,1.98,.57,.28,.42,1.95,1.05,1.82,520
2,12.33,1.1,2.28,16,101,2.05,1.09,.63,.41,3.27,1.25,1.67,680
2,12.64,1.36,2.02,16.8,100,2.02,1.41,.53,.62,5.75,.98,1.59,450
2,13.67,1.25,1.92,18,94,2.1,1.79,.32,.73,3.8,1.23,2.46,630
2,12.37,1.13,2.16,19,87,3.5,3.1,.19,1.87,4.45,1.22,2.87,420
2,12.17,1.45,2.53,19,104,1.89,1.75,.45,1.03,2.95,1.45,2.23,355
2,12.37,1.21,2.56,18.1,98,2.42,2.65,.37,2.08,4.6,1.19,2.3,678
2,13.11,1.01,1.7,15,78,2.98,3.18,.26,2.28,5.3,1.12,3.18,502
2,12.37,1.17,1.92,19.6,78,2.11,2,.27,1.04,4.68,1.12,3.48,510
2,13.34,.94,2.36,17,110,2.53,1.3,.55,.42,3.17,1.02,1.93,750

SUMMARY

The UCI (University of California, Irvine) data sets are the standard benchmark data sets used in the data mining research and practice for testing new algorithms [1, 3]. In order to bring some standard among the performance comparison of new algorithms against standard algorithms, it is mandatory that researchers prove against these data sets. They are freely available and downloadable from [2], in a variety of data formats. But the *arff* format is the favorite among the lot as it is the native format for WEKA package. About four data sets have been described with details in this chapter along with partial data.

EXERCISES

1. Describe the UCI data set named "LETTER," with its various details and sample data.
2. What is the significance of UCI data sets in data mining research?

3. Using WEKA, compare the performances of standard data mining algorithms such as decision trees and naïve Bayes classifier over the following data sets:
   a. Weather.arff b. Car.arff c. Diabetes.arff

REFERENCES

1. K.P. Soman, S. Diwakar, and V. Ajay, *Insight into Data Mining Theory and Practice*, PHI Publishers, 2006.
2. http://archive.ics.uci.edu/ml/datasets.html
3. R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

Scanned by CamScanner