

CSE5230 Algorithm Implementation Assignment

MONASH UNIVERSITY

Faculty of Information Technology

CSE5230 Data Mining

Semester 2, 2004

Due: Friday 24th September, 2004

1 Aim

The aim of this assignment is for you to demonstrate your detailed understanding of a simple data mining algorithm. You will do this by writing a program which implements the algorithm, and explaining how the code works to your tutor. You will also demonstrate the algorithm on a test data set provided by the lecturer.

2 Choice of Algorithms

You may choose one to implement any one of the follow algorithms:

- k -means clustering
- The Naïve Bayes classifier
- ID3 decision tree

Handouts explaining each of these algorithms will be available from the unit web site during week 3. Students interested in implementing other algorithms (*e.g.* Backpropagation Neural Network, Self-Organizing Map, *etc.*) should contact the lecturer before the end of week 3 to seek approval.

3 Platform for Implementation

You can implement the algorithm using the language and platform of your choice. The only constraint is that you must be able to demonstrate your code during tutorials. This means that you must use either a language available in the labs, or bring in a lap-top to do the demonstration.

4 Data Sets

Sample data sets will be provided on the unit web site for use when developing their algorithms. These datasets will be in ASCII text files. The first line of the file will specify the number of attributes p that

each data point has. The second line will specify the number of data points in the file, n . The third line will be a comma separate list of attribute names. Attribute names cannot contain whitespace. The fourth line will be a comma-separated list of data types, indicating the type of each attribute. The possible types are: `Nominal`, `Ordinal`, `Numerical`. The next p lines will specify possible values for each of the p attributes. Lines corresponding to numerical attributes will be blank: all values are considered possible *a priori*. For a nominal attribute, the line will contain a comma-separated list of all the possible values the attribute can take on. For an ordinal attribute, the line will contain an ordered, comma-separated list of all the values the attribute can take on. The following n lines will each specify a data point, as a comma-separated list of attribute values.

For example, consider a file containing:

```
4
3
Gender, Age, IncomeRange, Height
Nominal, Numerical, Ordinal, Numerical
M, F

Low, Medium, High

M, 34, Medium, 1.72
F 25, High, 1.6
M, 21, Low, 1.85
```

This file says that we are considering a data set where the data points have four attributes. There are three data points in the data set. The attributes are named `Gender`, `Age`, `IncomeRange`, and `Height`. `Gender` is a nominal attribute, `Age` is a numerical attribute, `IncomeRange` is an ordinal attribute, and `Height` is a numerical attribute. `Gender` can take on the values `M` and `F`. `Age` is numerical, so all values are possible. `IncomeRange` can take on the values `Low`, `Medium` and `High`, in that order. `Height` is numerical, so all values are possible. The three data points have the values (`M`, `34`, `Medium`, `1.72`), (`F`, `25`, `High`, `1.6`), and (`M`, `21`, `Low`, `1.85`).

You will have to consider how each attribute should be encoded before being provided as input to your algorithm. This will be discussed in the lecture on pre-processing for data mining.

5 Assessment

Assessment will be via the demonstration of your code to your tutor, using a newly provided test data set, and the explanation of the code to the tutor. The explanation of the code will be done by showing the source code on the computer screen and doing a “walk-through” with the tutor, explaining how each section works, and answering any questions the tutor might have. Students must demonstrate their understanding of the algorithms and data structures they have used.

Marks will be awarded as follows:

Correctness of Implementation:	10
Understanding demonstrated in walk-through:	10

Appointments for walk-throughs will be organized after week 10.