Australian Government
Department of Health and Ageing

Medicare Financing & Analysis Branch

Practice Categorisation

Model Methodology and Decision Trees

Medicare Financing and Analysis Branch
Department of Health and Ageing
March 2012

PROTECTED

Contents

PROTECTED

PROTECTED

1 Introduction

In this paper, we introduce practice categorisation methodologies for classifying both practice size and service patterns. Furthermore, we have developed a methodology to interpret and validate practice categories. Throughout a set of experiments, it is evident that our categorisation methodologies are able to classify practices effectively and efficiently.

In general practice analysis and monitoring, there are many applications for practice classifications. For example, when comparing operating behaviour between corporate and non-corporate practices, similar groups of corporate against non-corporate practices are needed so as to obtain unbiased analytic results. A categorisation mechanism which can classify the whole set of practices into several categories, according to their intrinsic ‘similarity’ is essential to this process. In this document, ‘similarity’ refers to either the profile of the practice itself or the way they provide services to patients.

The profile of a practice can be represented by a combination of factors, especially the number of work days, total number of services, total full-time service equivalent (FSE[1]), total schedule fee and total number of after-hour services. It indicates the capacity of each practice to provide primary care services. We refer to this practice profile as the size category.

Furthermore, we use service patterns to describe the way practices provide services to patients. Categorisation of practice services patterns indicates the characteristics of the services each practice provides. It is derived from the proportion of typical service groups, such as GP attendances and GP mental health services that the practice provides.

This paper is organised as follows: an overview of practice categorisation methodologies is provided in Section 2; experiments and results are discussed in Section 3, followed by a conclusion in Section 4. In addition, detailed background information and theories of related data mining techniques are included in the Appendix.

This work has been undertaken as part of the Department of Health and Ageing project, “Tracking the Effects of Corporate Practices on Medicare Outlays”. The general practice categorisation methodologies described in this paper have been developed for use in the project. Further experimental and exploratory work to refine and confirm the methods would be required before wider application could be considered.

---

1 Full-time service equivalent is an experimental workforce measure under development by Medicare Financing and Analysis Branch intended to be more suited to detailed statistical analyses particularly over a time series than the current summary measures, headcounts, full-time equivalent (FTE) and full-time workload equivalent (FWE).

PROTECTED

2 Categorisation Methodologies

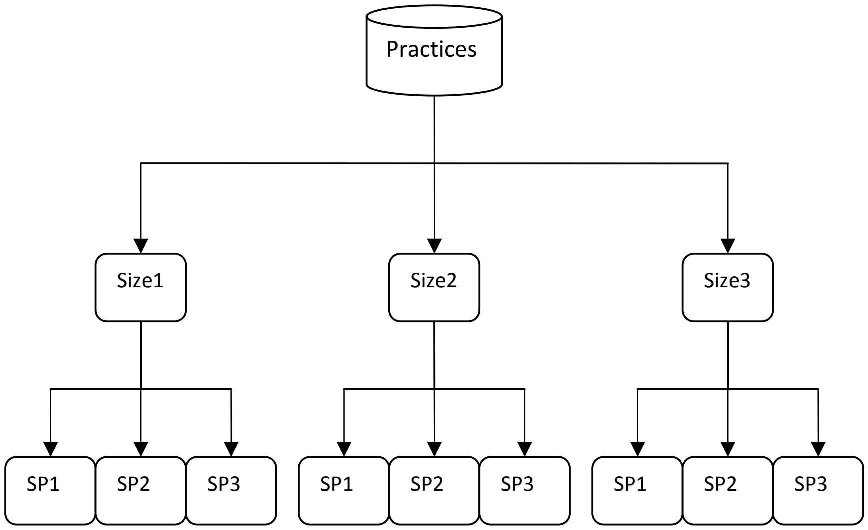
This section provides an overview of practice categorisation methodologies, and discusses how they can be applied to general practice. We briefly introduce relevant techniques used in categorisation. Further detail of the technical theories behind these methods is included in the Appendix.

2.1 Overview of categorisation

This paper provides methodologies to identify groupings of “like” practices in terms of practice size and service patterns by using data mining techniques. The two sets of categories are not allocated in parallel, they have a hierarchical structure.

Figure 1 demonstrates the final outcome that this methodology will achieve, with practices initially categorised by their size (Size1, Size2…) then further split into several service pattern categories (SP1, SP2). The size categories are determined by their size-relevant variables, while the service pattern categories are determined by a significantly high proportion of a specific service group.

Figure 1: Hierarchical structure of practice categorisation



As the exploratory variables for categorising practices are multi-dimensional and usually interact with each other, it is hard to perform classic statistical tests to identify and distinguish each category. Therefore, we will employ techniques in data mining to solve this complex statistical problem. The following Subsections (2.2, 2.3), introduce clustering, one of the most important data mining techniques, as our categorisation algorithm.

In order to interpret categorised results and set up corresponding decision rules for those categories, another popular data mining technique, decision trees, will be discussed in subsection 2.4.

PROTECTED

2.2 Clustering for size categorisation

Within data mining, clustering is perhaps one of the most important tools for both exploratory and confirmatory analysis. Clustering is a technique used to discern meaningful patterns in unlabeled data by grouping data points that are 'similar'.

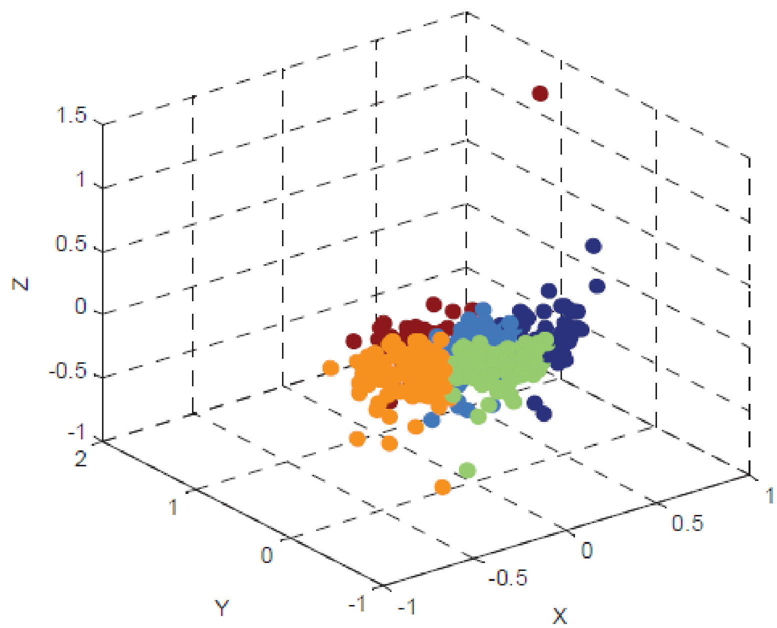
A cluster is a collection of data objects that are similar to other objects in the same cluster but are dissimilar to objects in other clusters. A comprehensive description of the theory and algorithms of clustering is included in Appendix 5.1.

We apply the clustering algorithm to our experimental practice data for categorisation. The relevant variables, such as work days, total number of services, total full-time service equivalent (FSE), total schedule fee and total number of after hour services, are selected by using our domain knowledge and input into clustering algorithm as the exploratory variables.

The objective of clustering is to identify similarities between each pair of practices, by using the input exploratory variables. The clustering algorithm is able to isolate practice groups from others with enough outer distance. In other words, it can represent statistically significant differences between each pair of groups.

Figure 2 demonstrates a 3-dimensional plot for clusters on general practice data (variables have been normalised) using the clustering algorithm. In Figure 2, it would be hard to identify separate groups within the original data by simply viewing the data graphically. However, the clustering algorithm is able to automatically identify groups according to their inner similarity and outer deviation, from the statistical point of view.

Figure 2: Visual presentation of clusters in 3-dimensional PIP data



PROTECTED

2.3 Clustering for service pattern categorisation

We use the same clustering technique to explore service pattern categories among practice data. As our target has been changed to the service pattern, we have to select another set of variables which are able to indicate characteristics of the services each practice provides. The selected exploratory variables are derived from the proportion of typical service groups, such as GP attendances and GP mental health services that the practice provides.

As our target for classification is a general practice, the statistical term, 'significantly high proportion', means more than 5% of the total services belonging to a specified service group other than GP attendances. For example, if more than 95% of services for a practice belong to the GP attendance group, the practice would be identified as a *Pure GP Practice*. If a practice has a significantly high proportion of GP mental health services (e.g. 10%), it would be regarded as a *GP Mental Health Practice*.

However, we should note that a high proportion of practices do not have any significantly outstanding service type. In other words, the services they provide are 'normally' distributed. Therefore, we allocate these practices to the normally distributed category. The service pattern categorisation experiments will be discussed in Section 3.3

2.4 Category interpretation

While clustering indicates suitable groupings, it does not of itself provide a method to readily classify new practices or to present the classifications in an understandable way. Therefore, after the exploration of clusters, it is necessary to perform further interpretation work. We selected another popular data mining technique, decision trees, to split the clusters from the original database by creating a set of decision rules. These decision rules are very explicit and easy to interpret, which in turn makes it simple to characterise the target category and convenient to implement the categorisation in future analysis.

Decision trees refer to a hierarchical model of decisions and their consequences. Decision trees are one of the most powerful directed data mining techniques as they can be used on a wide range of problems and they produce models that explain how they work. More details of the technique, including its theory and algorithms, are provided in Appendix 5.2.

In the category interpretation procedure, the practices in the desired category are labelled as a 'target', while the rest are labelled as 'non-target' practices. By using the decision tree algorithm, the most important variables to identify targets will be automatically selected. Meanwhile, the decision rules can be created by using the selected variables and the values used to distinguish between targets and non-targets.

Use of decision trees forms the second part of our methodology. It allows for the groupings discovered through clustering to be applied to new practices. Section 3.4 provides results for interpreting an explored size category that demonstrates the usefulness of the methodology in the case of general practice information.

PROTECTED

3 Experimental Results

3.1 Experiment settings

As discussed in Section 2.2 and Appendix 5.1, the clustering algorithm groups data instances according to their inner similarities. The algorithm is well suited to explore correlations among practices and find similar groupings. As all the attributes we are using are numeric, we can use the Euclidean distance to measure similarity in practice data.

We created a practice data base from practices enrolled in the Practice Incentive Program (PIP) in order to explore the data and inform the choices in the final practice categorisation model.

Aboriginal Medical Service practices (AMS) were excluded from the practice dataset, as they are known to behave differently to other general practices. Practices which were only open for a short period of the year (less than 200 calendar days) were also excluded as atypical. After these exclusions, 4,476 practices remained which, in combination with a set of exploratory variables, formed our experimental practice base data.

3.2 Experiments – size categorisation

Table 1 shows summary statistics for the clustering results. There are three clusters identified. Based on their statistical features, we can intuitively name them as the ‘Large’, ‘Medium’ and ‘Small’ practice categories.

Table 1: Practice size categorisation results on practice sample data

Category	No. of Practices	Proportion of Practices	Average FSE	Average No. of Providers	Average No. of Services	Average Work Days	Average After Hour Service Proportion
Large	413	9.2%	14.1	16.3	71,055	348	9.4%
Medium	2,621	58.6%	4.7	6.1	22,738	315	2.5%
Small	1,442	32.2%	2.0	2.6	9,406	256	0.7%

Based on the statistics in Table 1, we notice that there are significant differences among these size categories. Furthermore, we can identify the practice profile for each category. For example, the practices in large size category have 348 work days on average. It means that they are almost open all days in a year except some public holidays. For the medium size category, the average work days is down to 315 days which means the medium size practices are almost open 6 days a week except Sundays. And the last, small size practices are only open 5 days a week on average.

PROTECTED

3.3 Experiments - service pattern categorisation

In order to explore service patterns, we ran another set of experimental clustering tests on the same base data. As introduced in Subsection 2.3, the service pattern categorisation explores the statistically significant difference between service groups.

Table 2 shows the major statistical summary for identified service patterns. These service pattern categories are based on their significant service proportion characteristics. For example, in the ‘Pure GP Practice’ category, all practices share the same significant service type, ‘GP attendance services’, and their average service proportion of this type is about 95.8%.

These service pattern categories are able to explicitly indicate the service features of practices, and can be used to identify ‘similar’ practices (similar size and service pattern) in future practice micro-analysis.

The practices which are not located in these service pattern categories are those that do not have any significant service preference compared to other services and their services are ‘normally’ distributed.

Table 2: Practice service pattern categorisation results on practice data

Service Pattern Category	Service Proportion Characteristic	Number of Practices
Pure GP Practice	95.8%	885
GP Mental Health Practice	13.2%	59
GP Chronic Disease Practice	12.6%	169
Practice Nurse/AHW Practice	12.0%	634
Pathology Practice	12.6%	15
Operations Practice	7.8%	229
Obstetrics and Midwives Practice	8.7%	17
Normally distributed	Normally distributed	2468

We applied the service pattern clustering models on each size category. Table 3 shows the category distribution summaries.

The results show how the service pattern categories are distributed in those three size categories. For example, the first row demonstrates that the ‘Medium’ size category captures 50.7% of Pure GP practices while the ‘Small’ size category allocates 47.0% of them. However, the ‘Large’ size category only represents 2.3% of Pure GP practices.

PROTECTED

Table 3: Category distribution summaries under hierarchical structure

Service Pattern Category	Proportion in Large Size Category	Proportion in Medium Size Category	Proportion in Small Size Category
Pure GP Practice	2.3%	50.7%	47.0%
GP Mental Health Practice	0.0%	35.6%	64.4%
GP Chronic Disease Practice	2.4%	48.5%	49.1%
Practice Nurse/AHW Practice	9.3%	60.9%	29.8%
Pathology Practice	6.7%	20.0%	73.3%
Operations Practice	5.3%	48.0%	46.7%
Obstetrics and Midwives Practice	5.9%	52.9%	41.2%
Normally distributed	12.2%	62.4%	25.4%

Based on these distribution tables, the characteristics of both size and service pattern categories can be further explored. For instance, the GP Mental Health service pattern is more likely to appear in the small size practices, but is hard to find in the large size practices. In addition, across all significant service patterns, most of the large size practices can not be categorised as any of these service patterns. In other words, the large size practices usually have a mixture of service types without one of the service preferences identified.

3.4 Experiments - category interpretation

In order to verify the suitability of the theories described in this paper to categorise general practices, we ran an experiment applying the methodology to the “Small” size practice category.

In the pre-processing procedure, we assigned the label ‘1’ to practices in the small size category as the decision target, and the label ‘0’ to the rest of practices. The decision tree algorithm automatically selected a set of principal exploratory variables that effectively distinguished the target from all practices.

The results of this experiment are shown in Figure 3 (page 12). Three principal exploratory variables (work days, FSE and number of doctors) were selected. In other words, we can identify whether a practice belongs to the small size practice category by just using these three factors. Meanwhile, the decision rules are also clearly presented in the tree structure graph.

In Figure 3, the target leaves are marked in deep blue, the paths which link from root node to the target leaves are marked in bold. The decision rules are formed by these paths.

In this instance, the decision rules are:

> if the number of ***Working Days*** is less than 285.5 and
>
> the ***Total FSE*** is less than 4.55 and
>
> the ***Number of Doctors*** is less than 9.5 and
>
> the ***Total FSE*** < 3.89.

Note that these are experimental results, and may not concur with the final decision model produced if this methodology is applied.

PROTECTED

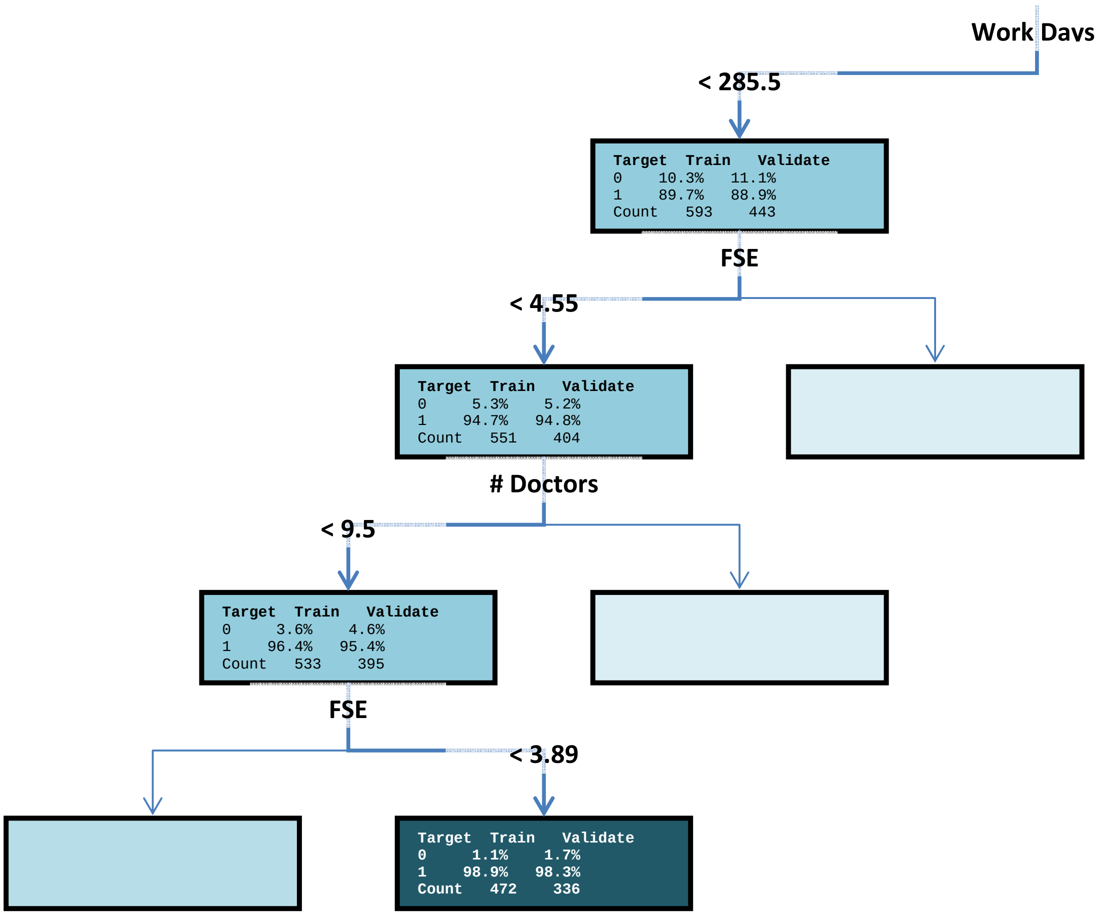
Then we gain an explicit and practical idea of what constitutes a small size practice by using the decision rules. Looking at the decision tree, the character of the small size practice category can be simply interpreted as a group of practices which:

- are open no more than five days a week, and
- have no more than four full-time doctors.

Comparing the results with the target values, we see that practices are identified by the decision tree as a small size practice with 98.3% accuracy compared to the clustering results. Although the remaining 1.7% differs from the clustering model, they fit with an easily understood real-world definition of “Small”.

This experiment validates the feasibility and suitability of the clustering and decision tree methodologies described in the paper for future work in categorizing general practices for analysis purposes.

Figure 3: A decision tree example for interpreting the small size practice category



PROTECTED

4 Conclusion

In this paper a framework for practice categorisation is introduced. The practices will be initially categorised using clustering by their size, which represents the multi-dimensional capacity each practice has for providing primary care services, determined by size-relevant exploratory variables such as

- Work days,
- Full-time service equivalent; and
- Number of doctors.

Then, the practices will be further split into service pattern categories. These service pattern categories are derived by the proportion of types of typical service groups, which indicates the characteristics of the service patterns they have provided.

Furthermore, this paper also introduces two kinds of data mining techniques used in categorising and interpreting practices.

Based on the experiments on a sample set of practice data, the integration of the two techniques is competent to categorise practices effectively and explicitly.

PROTECTED

5 Appendix

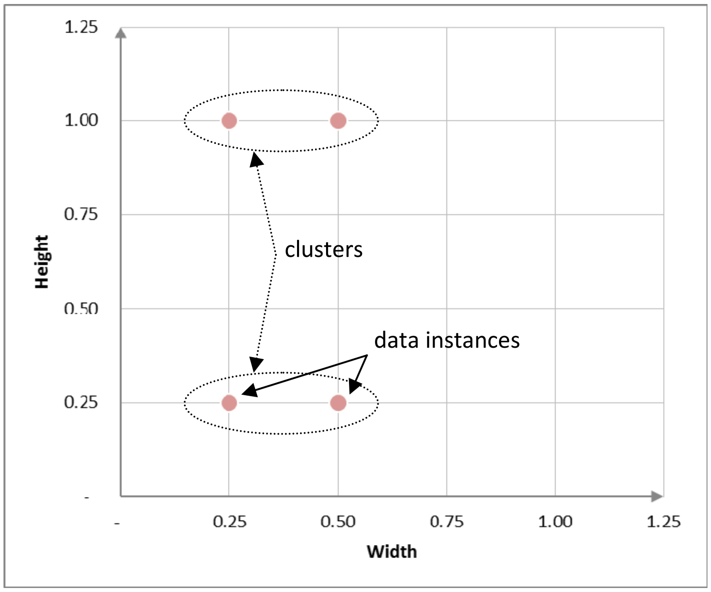
5.1 Clustering theory and methodology

Clustering theory

Clustering is perhaps one data mining's most important tools for both exploratory and confirmatory analysis (Cios, et.al: 1998). Clustering is a technique used to discern meaningful patterns in unlabeled data by grouping data points that are 'similar'. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.

We can show this with a simple graphical example:

Figure 4: Demonstration of a 2-dimensional cluster



In Figure 4we can easily identify two clusters; the similarity criterion is distance: two or more objects belong to the same cluster if they are 'close' according to a given distance.

The distance measure between data points is an important component of clustering algorithms. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean metric is sufficient to successfully group similar data instance (Agrawal, et.al: 1998).

The Euclidean distance between points p and q is calculated as:

$$dist(p,q) = \sqrt{(q_i - p_i)^2}$$

PROTECTED

Clustering algorithm

The most well-known and commonly used partitioning methods are "k-means". The k-means algorithm takes the input parameter, k , and partitions a set of n objects into k clusters so that the resulting intra-cluster similarity is high whereas the inter-cluster similarity is low. Clustering similarity is measured in regard to the mean value of the objects within the cluster. This can be viewed as the cluster's 'centre of gravity'.

The algorithms procedure is as follows:

> Firstly, it randomly selects k of the objects to represent the k cluster means or centres.
>
> Each of the remaining objects is then assigned to the cluster to which it is the most similar, based on the distance between the object and the cluster mean.
>
> It then computes the new mean for each cluster. This process iterates until the criterion function converges.

Typically the squared-error criterion is used, defined as:

$$E = \sum_{i=1}^{k} \sum_{r \in C_i} | x - m_i |^2 ,$$

Where x is the point space representing the given object, and m_i is the mean of cluster C_i (both x and m_i are multidimensional). This criterion tries to make the resulting k clusters as compact and as separate as possible.

The k-means algorithm procedures are summarised as follows (Hinneburg and Keim: 1998):

> **Algorithm 1 (k-means)**: The k-means algorithm for partitioning based on the mean value of the objects in the cluster**.**
>
> **Input**: The number of clusters k , and a database containing n objects**.**
>
> **Output**: A set of k clusters which minimises the squared-error criterion**.**
>
> **Method**: The k-means algorithm is implemented as follows:
>
> > Arbitrarily choose k objects as the initial cluster centre;
> >
> > **Repeat**
> >
> > > 1. (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
> > >
> > > 2. Update the cluster means, i.e., calculate the mean value of the objects for each cluster;
> >
> > **Until** no change.

Based on the algorithm described above, k-means is very efficient in processing large datasets because the computational complexity of the algorithm is:

$$O(nkt)$$

where

n is the total number of objects,

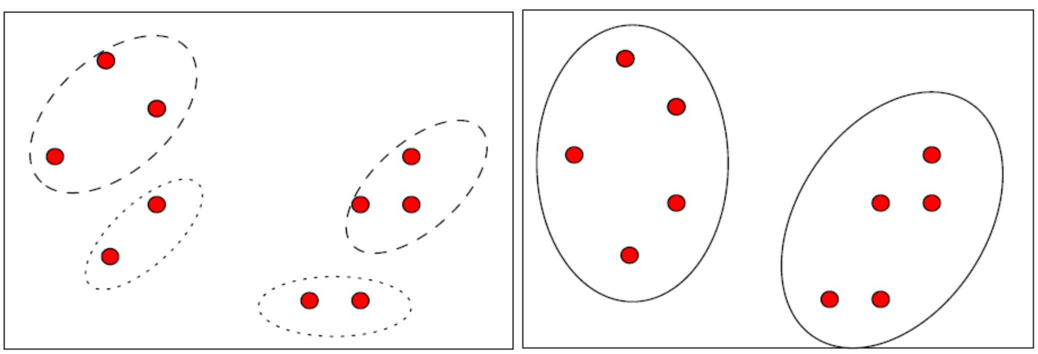
k is the number of clusters, and

t is the number of iterations.

Figure 5 (below) illustrates the k-means method. In Figure 5, there are a set of objects located in a rectangle. Suppose the user would like to cluster the objects into two clusters. According to the k-means algorithm above, we arbitrarily choose 2 objects as the initial two cluster centres. Then each object is distributed to the chosen cluster domains based on which cluster centre is the nearest. Such a distribution forms a silhouette encircled by the bold dashed curve, as shown in Figure 5(a).

This kind of grouping will update the cluster centres. That is the mean value of each cluster is recalculated based on the objects in the cluster. Relative to these new centres, objects are re-distributed to the chosen cluster domains based on which cluster centre is the nearest. Such a re-distribution forms a new silhouette encircled by a solid curve, as shown in Figure 5 (b).

Figure 5: Clustering of a set of points based on the k-means method



(a) (b)

The next section discusses use of the clustering algorithm to categorise general practices, and how the categorisation model works.

5.2 Decision tree theory and methodology

Decision tree theory

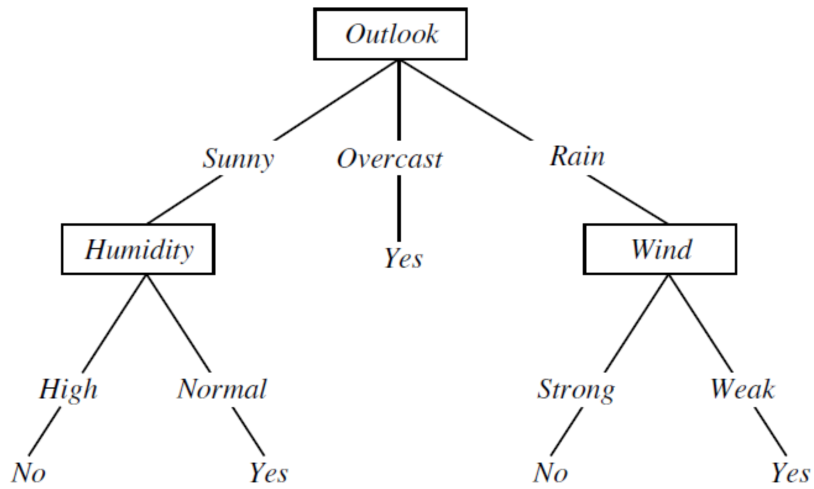
The remainder of this subsection introduces the basic framework of decision trees and how they are used in category interpretation and implementation.

Decision trees refer to a hierarchical model of decisions and their consequences. Decision trees are one of the most powerful directed data mining techniques as they can be used on a wide range of problems and they produce models that explain how they work (Cios, et.al: 1998).

PROTECTED

Figure 6shows a naïve decision tree example for deciding whether a tennis court is likely to be open. In this example, the weather outlook is split by the three major weather conditions, sunny, overcast and rain. One of the weather conditions (overcast) results a direct decision, meanwhile, the remaining conditions (sunny and raining) are further split by humidity and wind conditions separately. Based on this tree structure graph, a decision can be made whether to go to the tennis court according to the observed weather conditions.

Figure 6: A naïve decision tree for finding a tennis court open based on weather



From a mathematical point of view, a decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called a ‘root’ that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as an ‘internal’ or ‘test’ node. All other nodes are called ‘leaves’ (also known as ‘terminal’ or ‘decision’ nodes).

In the decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attribute values. In the simplest and most frequent case, each test considers a single attribute, and the instance space is partitioned according to the attribute’s value. In the case of numeric attributes, the condition refers to a range.

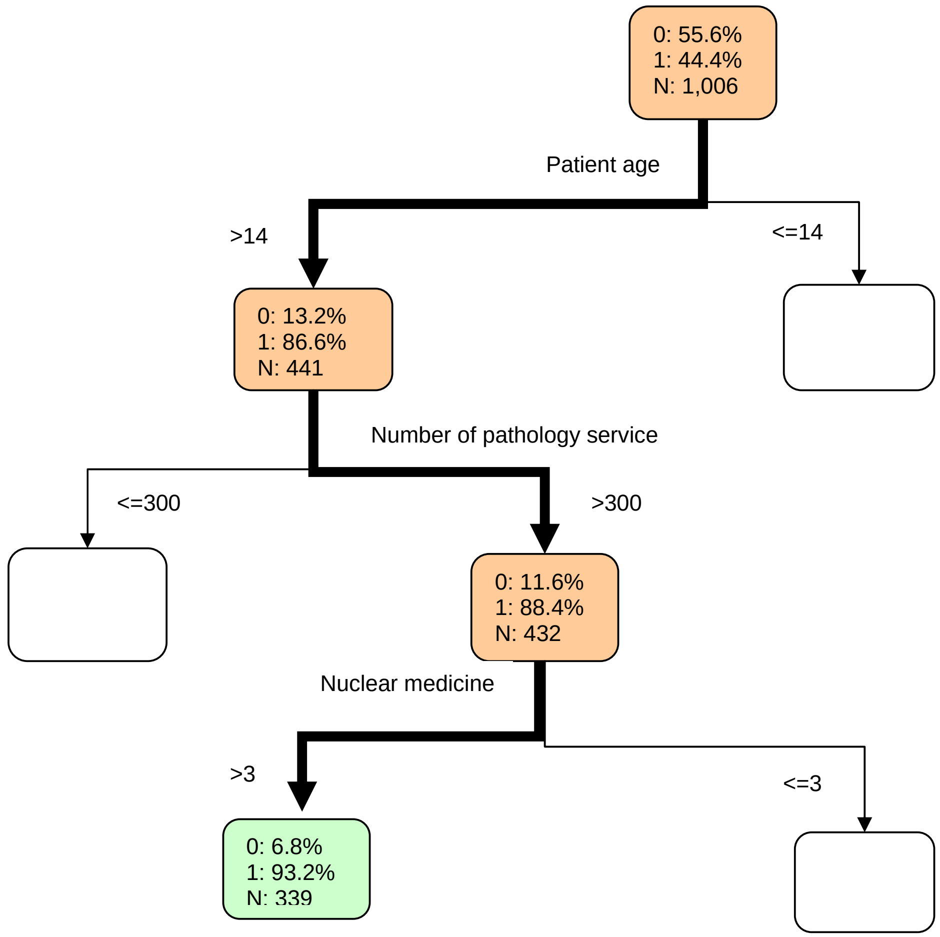
Each leaf is assigned to one class representing the most appropriate target value. Alternatively, the leaf may hold a probability vector (affinity vector) indicating the probability of the target attribute having a certain value.

Example of a decision tree

Figure 7 demonstrates a decision tree that explores identification of leukaemia patients by their treatment profile. The target leaf is marked in green and other leaves are left blank. All nodes and paths which link between root and the target leave are highlighted. Each node corresponds with a certain characteristic and the branches correspond with a range of values. These ranges of values must give a partition of the set of values of the given characteristics. For example, the root (base data) is initially separated by patient age (greater than 14 years old or not).

PROTECTED

Figure 7: Decision tree identifying leukaemia patients by their treatment profile



In this example, the decision tree demonstrates the treatment profile of leukaemia patients which is dominated by patient age, number of pathology services and number of nuclear medicine services. Furthermore, the target path with partition variables forms a set of rules which are helpful to identify other leukaemia patients outside the training database and monitoring their treatment program.

Instances are classified by navigating them from the root of the tree down to a leaf, according to the outcome of the tests along the path. Specifically, we start with a root of a tree; we consider the characteristic that corresponds to a root; and we define to which branch the observed value of the given characteristic corresponds. Then we consider the node in which the given branch appears. We repeat the same operations for each node encountered until we reach a leaf.

Creation of the decision tree

The algorithm for decision tree induction is shown at Algorithm 2 (Cios, et.al: 1998). It is a well-known "greedy" algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner.

The basic strategy is as follows:

- The tree starts as a single node representing the training samples (step 1).
- If the samples are all of the same class, then the node becomes a leaf and is labelled with that class (steps 2 and 3).
- Otherwise, the algorithm uses an entropy-based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes (step 6).
- This attribute becomes the 'test' or 'decision' attribute at the node (step 7). In this version of the algorithm, all attributes are categorical, i.e., discrete-valued. Continuous-valued attributes must be made discrete, by being expressed as a range.
- A branch is created for each known value of the test attribute and the samples are partitioned accordingly (step 8-10).
- The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered in any of the node's descendents (step 13).
- The recursive partitioning stops only when any one of the following conditions is true:
    1. All samples for a given node belong to the same class (step 2 and 3), or
    2. There are no remaining attributes on which the samples may be further partitioned (step 4). In this case, **majority voting** is employed (step 5). This involves converting the given node into a leaf and labelling it with the class in majority among samples. Alternatively, the class distribution of the node samples may be stored; or

There are no samples for the branch test-attribute= a_i (step 11). In this case, a leaf is created with the majority class in samples (step 12).

PROTECTED

Algorithm 2: Decision tree generation algorithm

Algorithm (Generate_decision_tree) Generate a decision tree from the given training data.

Input: The training samples, samples, represented by discrete-valued attributes; the set of candidate attributes, attribute-list.

Output: A decision tree.

Method:

1) Create a node N;
2) **if** samples are all of the same class, C **then**
3) return N as a leaf node labelled with the class C;
4) **if** attribute-list is empty **then**
5) return N as a leaf node labelled with the most common class in samples; (majority voting)
6) select test-attribute, the attribute among attribute-list with the highest information gain;
7) label node N with test-attribute;
8) **for each** known value a_i **of** test-attribute (partition the samples)
9) grow a branch from node N for the condition test-attribute= a_i;
10) let s_i be the set of samples in samples for which test-attribute= a_i; (a partition)
11) **if** s_i is empty **then**
12) attach a leaf labelled with the most common class is samples;
13) **else** attach the node returned by Generate_decision_tree (s_i, attribute-list – test-attribute);

PROTECTED

References

1. K. Cios, W. Pedrycz, and R. Swiniarski. *Data Mining Methods for Knowledge Discovery*. Boston: Kluwer Academic Publishers, 1998.

2. A. Hinneburg and D. A. Keim. *An efficient approach to clustering in large multimedia database with noise.* In Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98), pages 58 65, New York, NY, August 1998.

3. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. *Automatic subspace clustering of high dimensional data for data mining applications.* In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data, pages 94-105, Seattle, Washington, June 1998.

PROTECTED