Feature Selection for Temporal Health Records

Rohan A. Baxter, Graham J. Williams, and Hongxing He

CSIRO Mathematical and Information Sciences,
GPO Box 664, Canberra 2602, Australia,
{Rohan.Baxter,Graham.Williams,Hongxing.He}@cmis.csiro.au

Abstract. In this paper we consider three alternative feature vector representations of patient health records. The longitudinal (temporal), irregular character of patient episode history, an integral part of a health record, provides some challenges in applying data mining techniques. The present application involves episode history of monitoring services for elderly patients with diabetes. The application task was to examine patterns of monitoring services for patients. This was approached by clustering patients into groups receiving similar patterns of care and visualising the features devised to highlight interesting patterns of care.

1 Introduction

We are interested in the problem of clustering individuals given observed data about the individuals where the observed data does not naturally occur in vector form. Clustering algorithms are typically applied to data in vector form. For example, we may have k-measurements on a set of patients and so the measurements on each individual i are represented as a k-dimensional vector. For vector-form data well-known and widely-applied clustering techniques can be applied. Such techniques are generally model-based methods include mixture modelling [6], or distance-based methods [3].

Much real world data is actually in non-vector form consisting of observations of an individual, recording information at particular time points. Such variable-length event sequence data is described in Sect. 2, but examples include a patient's usage of medical services and an individual's stock trading behaviour. The data is characterised as irregular events where each event may encapsulate a different type of action.

The data mining practitioner wishing to cluster event sequence data appears to have three options. The first option is to convert the event sequence data into feature vectors [4]. A problem with this approach is that information is inevitably lost in the vectorisation process. The second option is to use a distance-based clustering method which allows for non-vector data. An edit-distance metric [5] which uses insert, delete and replace operations to turn one sequence into another is an example of this approach. A difficulty here is in defining an effective distance metric. A suitable distance metric needs to be created for each new application. The third option is the use of mixtures of a generative probabilistic model [2, 1]. This is an attractive approach but not further explored here.

We chose the first option for the application described in this paper. An aim was to minimise the loss of information relevant to the data mining objectives in choosing the feature vectors. We present three alternative feature vectors for representing medical event sequence data. Our exploration provides insights into the process of developing alternative feature sets. We identify feature sets that are useful for clustering event sequence data.

Sect. 2 describes the patient health record data and Sect. 3 describes the objectives for investigating patterns of care received by patients. Sect. 4 describes the feature vectors we have used in looking for patterns of care. To the best of our knowledge two of the three feature vectors we use here are novel. Clustering results and their visualisations are presented in Sect. 5.

2 Health Care Data

Medicare is the Australian Government's universal health care system. Each visit to a medical practitioner or hospital is covered by Medicare and recorded as a transaction in the Medicare Benefits Scheme (MBS) database. This data has been collected in Australia since the inception of Medicare in 1975. Such a massive collection of data provides an extremely rich resource that has not been fully utilised in the exploration of health care delivery in Australia.

For this current exploration we use a subset of de-identified data (to protect privacy) based on Medicare transactions from Western Australia (WA) for the period 1994 to 1998. Our particular focus is on patterns of care related to diabetes for elderly patients (over 65 years of age). We have only limited demographic information about each patient, such as age, gender and location. For each patient we also have the sequence of diabetes-related monitoring tests they have received over this time interval.

The four monitoring tests included in our dataset are given in Table 1. Glycated hemoglobin measurements (Gl) provide information about the accu-

Abbrev	Description	Guidelines
Gl	Quantitation of glycosylated hemoglobin.	2–4 times per year
Op	Ophthalmologic examination.	Every 1-2 years
Ch	Cholestorol measurement via lipid studies.	Every year
Al	Microalbuminuria test	Every year

Table 1. Types of services received by Patients and indicative guidelines.

mulated effect of glucose levels. Ophthalmologic examinations (Op) are important in the early identification of complications related to eye sight. Cholesterol measurements via lipid studies (Ch) help identify possible complications relating to heart conditions. Microalbuminuria tests (Al) provide early indications of possible future kidney function loss.

A sample patient record is illustrated in Fig. 1. The event sequence data can be augmented with any available vector based data.

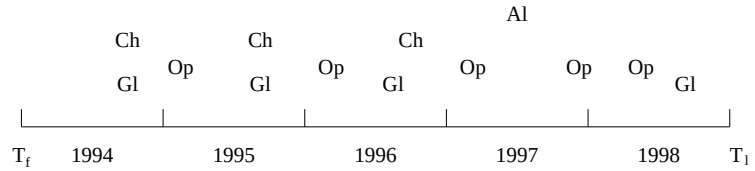

Fig. 1. A sample patient's health record, showing the four types of tests received over five years. The tests are: glycated hemoglobin (Gl); ophthalmology (Op); cholesterol (Ch); and micro-albuminuria (Al).

3 Patterns of Care in the Management of Diabetes

An important area in health population research is the investigation of patterns of care received by patients. Are there distinct patterns of care for these diabetes patients? Are there groups of patients receiving similar patterns of care? Are the patterns of care related to their doctor? Do patients of different age or gender or location receive differing patterns of care to other patients?

We have some prior expectations about the desired patterns of care for elderly patients with diabetes. The Australian National Health and Medical Research Council (NHMRC) publishes clinical guidelines for looking after patients. Patients with diabetes are at risk of developing complications such as eye problems, loss of kidney function and circulatory problems. The clinical guidelines recommend monitoring services, such as those in Table 1, be carried out at certain regular intervals. There is no compulsion for general practitioners to adhere to these guidelines and the guidelines cannot be expected to be appropriate for everyone.

To complicate matters, published clinical guidelines can differ in their details from state to state, and from country to country. We use the NHMRC guidelines as our starting point, but refer to other guidelines where they differ and where they may have an effect on clinical practice in Western Australia.

For example, according to NHMRC guidelines, glycated hemoglobin measurement should be done every six months (or every four months for some guidelines). Ophthalmologic examinations should be done every two years (or annually for some guidelines). The cholesterol measurement via lipid studies should be performed once a year. The microalbuminuria test should be done annually.

4 Selecting Features

We now present three methods for mapping the non-vector sequence data onto feature vectors. The first is the obvious *count* approach of having one feature for each type of service representing the number of times the service was used. The other two methods, which we call *average-residual-deviance* and the *gap*, are less obvious and overcome some shortcomings of the *count* method.

4.1 Count

In the *count* feature vector approach we have one feature for each type of service. Each feature contains the number of services received. The original sequence data, as shown in Fig. 1, is mapped to the features shown in Table 2. This

Patient	*Gl*	*Op*	*Ch*	*Al*
1	4	5	3	1
2	5	0	0	0
2	1	1	1	1
3	16	20	17	17

Table 2. Count Feature Vectors

feature representation has the advantage of being easily interpretable. However the obvious loss of information is a concern for the goals of our project. We have lost information relating to the time between successive services and also to the overall coverage of the services across the five years. For example, an individual with a count of 15 for a service appears to be well-monitored, but if those 15 services all occurred in 1994 and none occurred in 1995, 1996, 1997 and 1998, then that is a pattern we would like to identify.

4.2 Average, Residual, Deviance

We have devised the *average, residual, deviance* feature vector for capturing the required temporal information missed by the *count feature vector* approach.

Let $t^k_{i,j}(i = 1, 2, ...n_j)$ be the date of the ith service for service type j on patient k and n_j be the total number of type j services received by patient k. Define T_f and T_l to be the beginning and ending dates of the time interval covered by the study.

Definition 1 (Mean Interval). *The Mean Interval, $MI_{j,k}$, for the patient k on service j when $n_j > 0$ is defined as :*

$$MI_{j,k} = \frac{\sum_{i=1}^{n_j-1}(t^k_{i+1,j} - t^k_{i,j})}{n_j - 1} \tag{1}$$

If $n_j = 0$ then $MI_{j,k} = T_f - T_l$.

For example, we can calculate the Mean Interval for a patient who had thirteen tests for Quantitation of glycosylated hemoglobin on the following dates:

$$9044, 9272, 9377, 9527, 9592, 9766, 9875, 9985, 10101, 10154, 10334, 10413, 10510 \tag{2}$$

where, for computational convenience, these dates are expressed as the number of days since January 1st, 1970. The interval (in days) between two consecutive tests are then

$$228, 105, 150, 65, 174, 109, 110, 116, 53, 180, 79, 97 \tag{3}$$

For this patient we have $MI_{j,k} = \frac{228+105+150+65+...+116+53+180+79+97}{12} = 122.2$.

Definition 2 (Deviation Interval). *The Deviation Interval,$DI_{j,k}$, for patient k on service j for $n_j > 0$ is defined as:*

$$DI_{j,k} = \sqrt{\frac{\sum_{i=1}^{n_j-1}(t^k_{i+1,j} - t^k_{i,j} - MI_{j,k})^2}{n_j - 1}} \tag{4}$$

If $n_j = 0$ then $DI_{j,k} = T_f - T_l$.

For example, using the patient with the health record for a single test given in Eqn. (2), the Deviation Interval is

$$DI_{j,k} = \sqrt{\frac{(228 - 122.2)^2 + (105 - 122.2)^2 + (150 - 122.2)^2 + ...}{12}} = 47.9 \tag{5}$$

Definition 3 (Residual Time). *The Residual Time, $RT_{j,k}$, for patient k on service j is defined as:*

$$DI_{j,k} = t^k_{1,j} - T_f + T_l - t^k_{n_j,j} \tag{6}$$

For example, using the patient with the health record given in Eqn. (2), we have $T_f = 8765$ (January 1st, 1994) and $T_l = 10592$ (December 31st, 1997), so that the Residual Time is $DI_{j,k} = 9044 - 8765 + 10592 - 10510 = 361$

The feature *Mean Interval* measures the average interval between receiving the same service. The *Interval Deviation* measures whether the service intervals are regular or irregular. The third feature provides a way of accounting for the windowing effects of having data for 5 years only. The time interval from the window boundary to the time of the first service and from the last service to the window boundary are not considered in the definition of the first two feature definitions. The third feature is used to account for these boundary effects.

The feature vector for a service should have reasonably small values for all three features if the patient is treated according to the clinical guidelines. Typically, some patients do need more frequent services as their diabetic condition is serious. We do not consider the possibility of over-servicing by medical practitioners, where more services than are clinically necessary are provided.

Patterns of care contrary to the clinical guidelines can arise from insufficient numbers of services provided over the five years. This type of pattern is detected by the count feature vector and by a large Mean Interval value. The *average, residual, deviance* feature vector also represents patterns of care where the services provided are clustered in time, or are absent near the boundaries of the time window.

The features are still relatively easy to interpret. However, we now need twelve features in our present application instead of the four for the *count* feature vector.

4.3 Gap

Our third feature vector representation is the most specific to the task of investigating service patterns with reference to service clinical guidelines. The motivation is to describe the total length of time when the regular required tests are not carried out.

Once again, let $t^k_{i,j}(i = 1, 2, ...n_j)$ be the date of the ith service for service type j on patient k and n_j be the total number of type j services received by patient k. Define T_f and T_l to be the beginning and ending dates of the time interval covered by the study.

We require that service type j have a *desirable gap*, DG_j, as given by some clinical guidelines.

Definition 4 (Gap). *If patient k has $n_j = 0$ (the patient has no services) then the gap, $G_{j,k}$, is defined as:*

$$G_{j,k} = T_l - T_f - DG_j \tag{7}$$

If patient k has $n_j > 0$ (the patient has one or more services)

$$G^{initial}_{j,k} = \begin{cases} t^k_{1,j} - T_f - DG_j & \text{if } t^k_{1,j} - T_f > 0 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The following counts the time intervals between services received:

$$G^i_{j,k} = \begin{cases} t^k_{i+1,j} - t^k_{i,j} - DG_j & \text{if } t^k_{i+1,j} - t^k_{i,j} > DG_j \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

We then include the final service interval:

$$G_{final} = \begin{cases} T_l - t^k_{n_j,j} - DG_j & \text{if } T_l - t^k_{n_j,j} > DG_j \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The three Gap sub-parts are now summed:

$$G_{j,k} = G^{initial}_{j,k} + \sum_{i=1}^{n_j - 1} G^i_{j,k} + G^{final}_{j,k} \tag{11}$$

For example, using the patient with the health record given in Eqn. (2), and assuming that the $DG_1 = 120$ for the Quantitation of glycosylated hemoglobin test ($j = 1$). As shown previously, $T_f = 8765$ and $T_l = 10592$. The Gap between T_f and the first test is $9044 - 8765 = 179$ which is greater than DG_1, so it contributes $179 - 120 = 59$ to the sum. There are four time intervals exceeding DG_1 among the 12 time intevals. They are $228, 150, 174$ and 180 days respectively. Their contribution to the sum is $108, 30, 54$ and 60 days respectively. The last test was done on day 10510 and the gap between T_l and the last test is $10592 - 10510 = 82$, which is less than DG_1 and therefore contributes nothing to the sum. Therefore we have $G_{j,k} = 59 + 108 + 30 + 54 + 60 = 311$.

The advantage of this feature vector is that it is low-dimensional and easy to interpret. This feature is particularly useful when there is an expectation of regularity in the events and this regularity is to be explored.

5 Results

We used a model-based clustering program called Snob[7, 8] using a Bayesian mixture-modelling method with a Poisson distribution for the *count* feature vectors and a log-normal distribution for the *average, residual, deviance* feature vectors.

5.1 Clustering using *Count*

Fig. 2 gives the means and membership size of the 23 clusters found using Poisson mixture models. The Poisson distribution was suitable for these features because the counts are positive integers. Note that the counts are only approximately Poissson, because very large counts of services do not occur at all in practice. We now interpret these clusters. First recall that to receive care conforming to clinical guidelines for the *Gl* test over five years you would need between 10 and 15 tests. The population mean is 5 tests. It is apparent from Fig. 2 that most individuals do not receive conforming care because the large membership clusters (e.g., 4, 5, 6, 7) have mean counts below 6. Only two clusters (e.g., 2 and 20) have means of 10 or more. Cluster 2 individuals do not conform on the other two tests because they have means less than 3 for *Op* and *Cl*. In contrast, Cluster 20 individuals receive better than conforming care for all three tests. The 70 individuals in that group are apparently better looked after than all the others. The next best groups for all three tests are clusters 15, 16 and 17.

In follow-up work we plan to examine the characteristics (e.g., number of GP consultations, whether they are in the community or a nursing home, number of days in hospital) of individuals in the very 'good' and very 'bad' clusters to see if they can be distinguished from those receiving other patterns of care.

5.2 Clustering using *Mean, Residual, Deviance*

Fig. 3 gives the means and standard deviations of the 17 clusters found using log-normal mixture models. The log-normal distribution was suitable for these features because the mean, residual and standard deviation have positive continuous values. A mean interval of six months or so is indicative of care conforming to clinical guidelines for the *Gl* test. Looking at Fig. 3 we see that clusters 3, 6, 7, 8, 9, 10, 15 and 17 have mean intervals less than 10. Cluster 17 has less than 10 members and so we ignore it for the moment. Individuals in clusters 8 and 9 receive the best patterns of care for this population. The 700 cluster 3 individuals receive regular conforming *Cl* tests, but very infrequent *Op* tests. In follow-up work we hope to characterise these individuals further. It may be possible to devise a policy to improve their quality of ophthalmology care. Opposite to cluster 3, clusters 6, 7 and 9 receive frequent *Op* tests, but infrequent *Cl* tests.

At the other end of the quality of care, the 850 individuals in cluster 16 receive less care than the other individuals in the population.

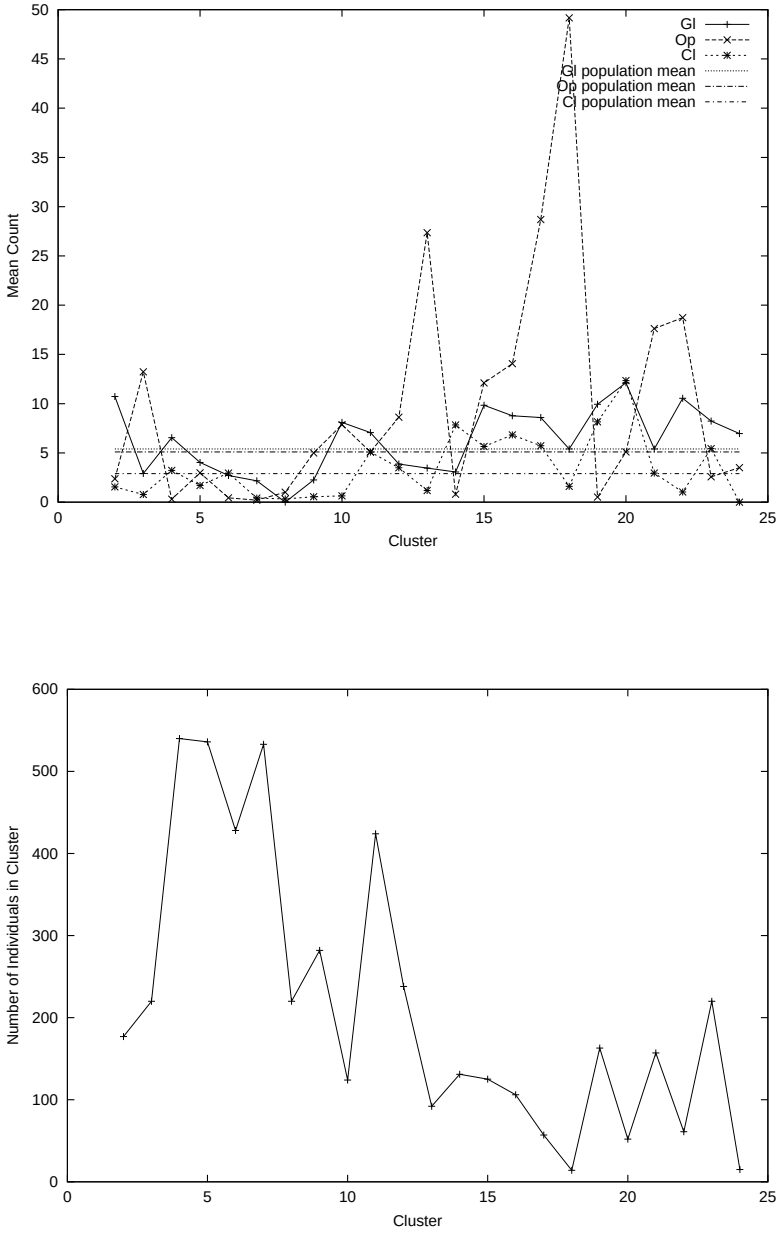

Fig. 2. Clustering results based on the *count* features.

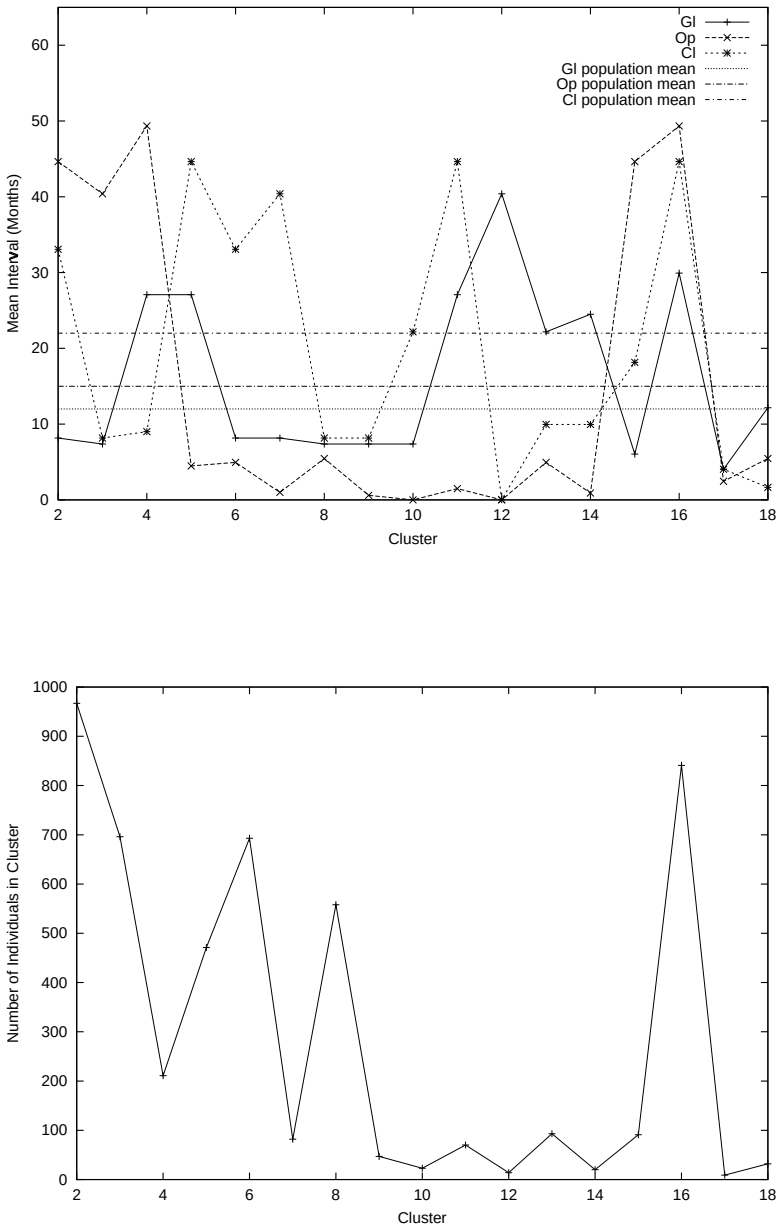

Fig. 3. Clustering based on /textitaverage, residual, deviance features. Residual and Deviance features were used in the clustering but are not shown.

We now compare the clustering results from Fig. 3 with those from Fig. 2 using a confusion matrix. The ith row of the confusion matrix contains the members of cluster i using *Count*. The individuals of cluster i are placed in column j if they belong to cluster j using *Average, Residual, Deviance*. If the two clustering approaches were identical, then one would expect the confusion matrix to contain one non-zero entry in each row and column. If the two clustering methods are independent, then one would expect a relatively uniform distribution of non-zero entries.

The confusion matrix is shown in Table 3. We see that there are indeed many zero entries indicating that the two clustering approaches result in related, but not identical, results. Intuitively, we would expect this if very high residual values are rare, thus making the count feature values highly correlated with the average feature values. The most interesting feature is that we consistently observe that individuals from a count cluster are distributed among one, two or three average, residual, deviance clusters. On closer examination, the average, residual, deviance clusters have similar mean average and deviance values, but differ in their residual value. This shows the value of using the residual feature to identify intensive patterns of care during a relative short time interval.

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	67	1	0	0	18	15	0	0	2	0	0	0	0	0
3	0	0	0	137	51	0	0	0	0	0	0	0	0	0
4	257	183	10	0	0	0	0	0	0	0	0	0	0	0
5	114	0	4	38	38	25	0	0	4	22	3	2	3	8
6	21	0	107	0	0	0	0	0	0	0	1	0	3	27
7	37	0	0	0	0	0	0	0	0	0	0	0	0	13
8	109	0	0	0	0	5	0	0	4	0	0	0	0	0
9	0	0	0	147	10	2	0	0	0	21	0	0	0	1
10	0	0	0	0	80	1	0	0	0	0	0	0	0	0
11	1	77	2	0	24	1	179	21	0	0	0	9	2	0
12	0	0	0	23	36	0	29	0	0	0	0	52	2	0
13	0	0	0	42	37	0	0	0	0	0	0	1	0	0
14	0	27	63	0	0	0	0	0	1	0	0	0	1	0
15	0	0	0	0	17	0	74	0	0	0	0	0	0	0

Table 3. Confusion Matrix. The rows show individuals from a cluster using *Count* distributed among the clusters using *Average, Residual, Deviance*. NB: Clusters 16–23 for *Count* and clusters 16–17 for *Average, Residual, Deviance* have been omitted.

5.3 Visualising the *Gap*

Fig. 4 provides a visualisation of the *Gap* for three different desired clinical guideline intervals (DG). The second of the three visualisations presents of Fig. 4 sets $DG = 6$ months. The distinct mode at zero indicates good conformance with the guidelines. In all three visualisations there is a mode around 20–24 months, worthy of further investigation: Is there some structural feature in the health system that has patients receiving this test every two years, rather than not at all (in the worst case). In general we note a peak at $Gap = 0$ and another

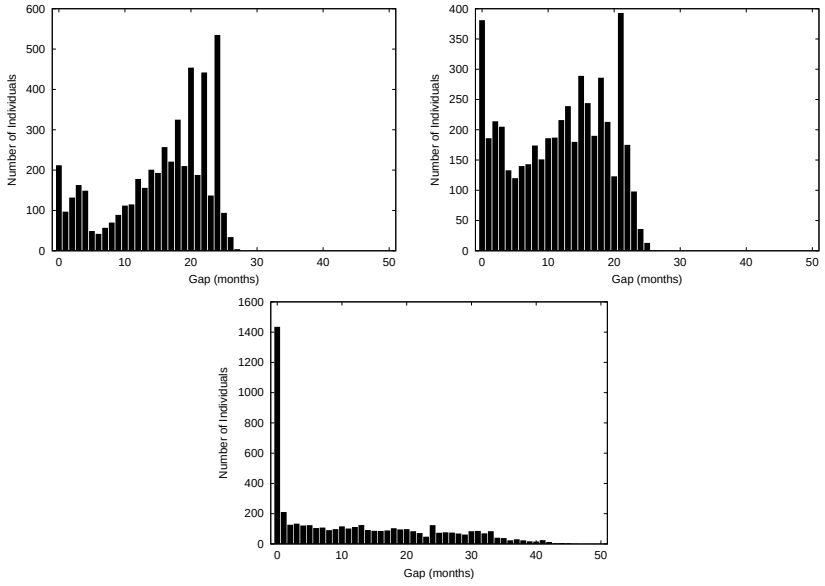

Fig. 4. Visualisation of *Gl* Gap with $DG = 3, 6$ *and* 12 months from left to right.

peak at the other end of the scale. These two peaks represent extremes: the first peak corresponds to conformance while the other peak corresponds to non-conformance to the guidelines.

Fig. 5 and Fig. 6 provide a visualisation of the *Op* and *Cl* tests for two different published clinical guideline intervals. At the right extreme are the individuals who did not receive any tests and so do not conform to the guidelines. At the left extreme are those individuals who conform to the guidelines. In between we see how patterns of care slowly degrade in terms of conformity. Note the mode at 12 months on the left-hand side panels (where $DG = 12$ months). This is indicative of the population of individuals receiving a precisely conforming pattern of care.

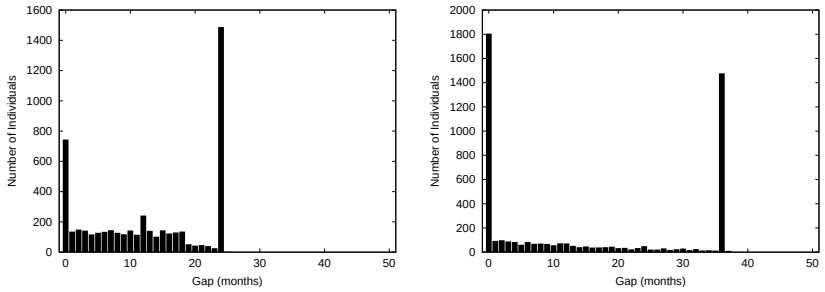

Fig. 5. Visualisation of *Op* Gap with $DG = 12$ *and* 24 months from left to right.

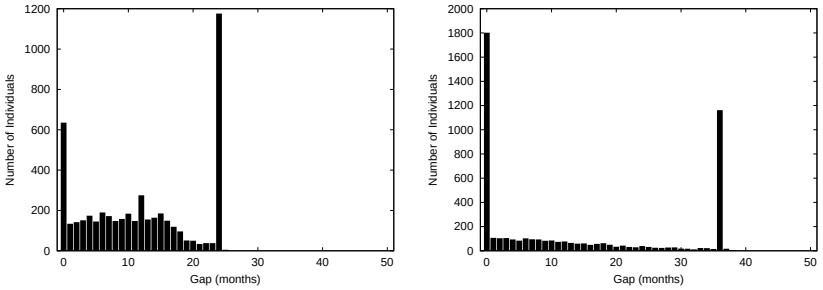

Fig. 6. Visualisation of *Cl* Gap with $DG = 12$ *and* 24 months from left to right.

6 Conclusion

We have considered three alternative feature vectors for representing variable-length patient health records. The feature vector of counts is the simplest, but can be misleading since it does not capture the distribution of patient care throughout the data window. The average, residual, variance feature vector overcomes this problem. For the specific task of characterising relationships to clinical guidelines, the gap feature vector most directly represents the required information. We expect the features created here for event sequence data for this health application will be applicable to other event sequence data such as trading and web log data.

References

[1] E. Arjas, H. Mannila, M. Salmenkivi, R. Suramo, and H Toivonen. Bass: Bayesian analyzer of event sequences. In *Proceedings in Computational Statistics (COMPSTAT'96)*, pages 199–204. Barcelona, Spain, Physica-Verlag, 1996.

[2] I. Cadez and P. Smyth. Probabilistic clustering using hierarchical models. Technical Report 99-16, Department of Information and Computer Science, University of California, Irvine, March 1999.

[3] A. Jain and R. Dubes. *Algorithms for Clustering.* Prentice-Hall, Englewood Cliffs, NJ, 1988.

[4] Huan Liu and Hiroshi Motoda. *Feature Selection for knowledge discovery and data mining.* Kluwer, 1998.

[5] P. Moen. *Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining.* PhD thesis, Dept. of Computer Science, University of Helsinki, Finland, 2000.

[6] J.J. Oliver, Baxter R.A., and Wallace C.S. Unsupervised Learning using MML. In *Machine Learning: Proceedings of the Thirteenth International Conference (ICML 96)*, pages 364–372. Morgan Kaufmann Publishers, San Francisco, CA, 1996.

[7] C.S. Wallace and D.M. Boulton. An information measure for classification. *Computer Journal*, 11(2):195–209, 1968.

[8] C.S. Wallace and D.L. Dowe. MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. *Statistics and Computing*, 10:73–83, 2000.