Surgical Length of Stay Classification

Elia El-DARZI and Revlin ABBI

School of Computer Science,
University of Westminster,
London, HA1 3TP, UK
eldarze@westminster.ac.uk

Abstract. Health care facilities operate administrative information systems which contain the admission and discharge dates of patient spells, used to obtain the length of stay (LOS) a patient has stayed in the health care facility. Understanding the different groups of patients with regards to their LOS and predicting LOS at admission would assist in making more informed and timely decisions on managing patients' care. In this paper we introduce a classification approach to distribute data, i.e. patient spells, into LOS classes (categories) of similar type.
Keywords: length of stay (LOS), classification, Gaussian mixture model
Math. Subject Classification: 62H30; 65C20

1 Introduction

Health care facilities operate administrative information systems to collect information on patient activity. Amongst other variables, the admission and discharge dates of patient spells are commonly recorded, and can be used to obtain the length of time a patient has stayed in the health care facility, referred to here as patient length of stay (LOS). LOS is often used as a proxy measure of a patient's resource consumption due to the practical difficulties of directly measuring resource consumption and the easiness of calculating LOS (Marshall A.H et al., 2005). Understanding the different groups of patients with regards to their LOS and predicting LOS at admission would assist hospital management and health professionals in making more informed and timely decisions on managing patients' care and planning for their discharge, and on allocating hospital resources (Marshall A.H et al., 2001).

In this paper we introduce a classification approach to distribute data, i.e. patient spells, into LOS classes (categories) of similar type. Classification is an approach often used to enhance understanding and allows predictions to be made in the presence of large volumes of historical data (Harper P.R, 2005). Based on previous studies in which this methodology has been applied (Abbi R et al., 2008, Abbi R et al., 2007b), we found that the classification model is often heavily influenced by the shorter stay classes. In such cases, the longer stay patient classes exhibit very low prediction accuracies. As such, we introduce a further processing step whereby a sensitivity analysis is performed to refine the LOS classes in order to increase the prediction accuracy of the derived classifier.

2 Methods

Based on random sampling (Cochran W, 1977), input data are split into two subsets of data. Two thirds of the data are used for training, and one third for testing. The GMM fitted to the LOS training data is a probability density model comprising of m Gaussian functions (Titterington D.M et al., 1985, McLachlan G.J and Peel D, 2000). Each Gaussian (component) j is described using three parameters, mean , variance and mixing coefficient . We limit the number of components of the GMM to eight, to avoid overfitting and to take account of human comprehension considerations (Miller G.A, 1956). The parameters of the model are estimated from the training data using the EM algorithm (Dempster A.P et al., 1977), implemented on the MATLAB technical software platform (Nabney I.T, 2004).

In order to find the optimum number of components we employ the Minimum Description Length (MDL) criterion (Rissanen J, 1978). Although MDL has shown to be effective for model selection (Walter M et al., 2001), it is also known for over-estimating the number of components (Walter M, 2002). As such, we assess the contribution of additional components based on the percentage decrease of the MDL value. The MDL criterion has also been validated against other commonly used criteria (Abbi R et al., 2007a), such as the Akaike information criterion (Akaike H, 1973) and the Bayesian information criterion (Schwarz G, 1978), and was found to suggest the same number of components.

To further aid the selection of an appropriate GMM, ten random samples of synthetic data are generated based on the parameters of each GMM. The 10^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th}, 99^{th}, 99.5^{th}, and 100^{th} percentile values for all samples are averaged, and used to make comparisons, measuring how well each GMM fits the LOS data.

Based on the selected GMM, we define the LOS classification scheme as a set of consecutive mutually exclusive LOS intervals defined according to the highest posterior probability *p(j—x)* of a LOS observation "belonging" to a component of the GMM, Equation 1. The posterior probability is derived using the Bayes rule, incorporating the conditional probability *p(x—j)*, Equation 2, prior probability *P(j)*, and the unconditional probability *p(x)*, Equation 3. The prior probability *P(j)* is obtained from the mixing coefficient ω_j, representing the prior knowledge of the proportion of group j to the overall population.

$$P(j|x) = \frac{p(x|j)P(j)}{p(x)} \quad (1)$$

$$P(x|j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} exp(\frac{(x-\mu_j)^2}{2\sigma_j^2}) \quad (2)$$

$$p(x) = \sum_{j=1}^{m} P(j)p(x|j) \quad (3)$$

Once LOS classes are defined, a decision tree is built from the training data using the C4.5 algorithm, programmed in the C programming environment (Quinlan J.R, 1993). In this case, the C4.5 algorithm attempts to find

a mapping between the patient records and the classes of the classification scheme.

Decision trees have shown to be particularly robust in healthcare applications compared with neural networks, regression models, and discriminate analysis (Harper P.R, 2005). They are computationally efficient and are naturally capable of handling datasets consisting of mixed data types (i.e. numerical and categorical variables). This is of particular value for the healthcare domain as health administrative information systems commonly exhibit both numerical (e.g. age) and categorical variables (e.g. diagnosis). In addition to their efficiency and ability to handle a mixture of different data types, decision trees are easily interpretable by humans and provide a clear indication of which variables are of most importance. The models can be easily converted into a set of logical rules which further aid their interpretability. The interpretable nature of decision trees enables the health professional to understand prospective predictions, i.e. take into account the variables and specific characteristics that form the basis of the predictions made. As such, justifications may be made, as the rationale of the prediction is known, thus enabling the health professional to approve or disapprove the prediction. A black-box approach like the neural network model would make it difficult to derive such justifications and thus confidence in the model would be at stake. Moreover, the average error rates of decision tree algorithms, statistical and neural network approaches are very similar (2000). In fact, amongst the various decision tree algorithms evaluated the C4.5 and CART algorithms tend to perform quite well. In addition, although neural networks perform slightly better in terms of prediction, the explanation capability that exists for decision trees is an important advantage (Perner P et al., 2001).

We evaluated the performance of the algorithm by using the overall and class accuracy, calculated as a percentage of correctly classified patient spells (Equation 4) and correctly classified spells for a given class j (Equation 5) respectively. In this case, S is the total number of correctly classified spells, and S_j is the total number of correctly classified spells for class j.

$$Overall\ Accuracy\ =\ \frac{S}{N}\cdot 100 \qquad (4)$$

$$Overall\ Accuracy_j\ =\ \frac{S_j}{N_j}\cdot 100 \qquad (5)$$

Sensitivity analysis

We further evaluate the tree using the testing data and by generating a confusion matrix (Han J and Kamber M, 2006). The confusion matrix is a common way to evaluate a classifier's ability to discern between spells belonging to different LOS classes within the classification scheme and is based on the various predictions made by the decision tree on the testing data.

A confusion matrix is an m by m matrix that contains information about the actual LOS class as well as the predicted LOS class of patient spells, where m is the number of classes within the derived classification scheme. Column j of the matrix represents the spells predicted as belonging to LOS class j, while row j represents the spells that actually belong to LOS class j. Correctly

predicted spells are therefore found in the diagonal of the matrix, i.e. *c(j,j)* within the matrix represents the correctly classified spells for LOS class *j*. If a patient spell belonging to class j is incorrectly classified as belonging to class *j+1*, then the matrix *c(j,j+1)* would be incremented by one, thus indicating that a spell belonging to class j was incorrectly classified to class *j+1*.

To improve the prediction accuracy of the classification model, we use the confusion matrix to provide insight of the influence that the classes within the classification scheme. For instance, for a given class *j*, the highest number within the confusion matrix in row j should be *c(j,j)*. If *c(j,j)* is not the highest, we then find the *c(j,h)* which consists of the highest number, amongst *c(j,q)*, where *q = 1,..., m*, and $q \neq j$. In predicting LOS, due to the skewed distribution, it is likely that class *h*, is a short stay class, influencing the classifier because the majority of patients stay short periods of time. This issue was previously referred to as the unbalanced class problem.

The objective is to refine the LOS classification scheme and to improve the ability of differentiating between patients belonging to different LOS classes. The procedure for conducting the sensitivity analysis and refinement of the classification model is as follows. Given a *m*m* confusion matrix, the objective is to maximise the *c(j,j)*, i.e. the number of correctly predicted spells for class *j*.

For notational purposes, the interval range for class j within a given classification scheme is denoted as j_{min} and j_{max}. For each class *j*, if *c(j,j)* within the confusion matrix is lower than *c(j,q)*, then we find the *c(j,h)* which consists of the highest number amongst cells *q*. We then decrease the LOS interval range for class *h* by one day, whilst increasing the range for the adjacent class by one day. More formally, if $c(j,j) \leq c(j,q)$, then we find *c(j,h)*. Moreover, if $h > j$, then we reduce h_{min} by one day, and increase $h - 1_{max}$ by one day. However, if $j¿h$, then we reduce h_{max} by one day and increase $h + 1_{min}$. In this case, *h-1* or *h+1* maybe equal to *j*.

Moreover, if any class j within the classification scheme consists of a very small percentage of spells, such that the prediction accuracy is very low, then we merge it with the adjacent class i.e. *j-1*.

Dataset

The data used is a surgical administrative discharge dataset. It consists of 7,723 records detailing the spells of patients undergoing some form of surgery in a tertiary hospital in Adelaide, Australia, discharged between 1^{st} July 1997 and 30^{th} June 1998. The variables describing each patient spell include the dates of hospital admission and discharge, the LOS in days, the gender of the patient, whether the patient was treated in a public or private hospital, whether the patient was admitted as an emergency case, and finally the diagnosis information - coded using major diagnostic categories (MDC). The MDC coding system consists of 25 categories, each of which corresponds to a single organ system. Each MDC for a patient is determined by the primary disease or condition for which a patient is hospitalised or treated.

3 Results

Before the proposed methodology is applied, all LOS observations are incremented by one, ensuring that the workload during the day is considered within the model (Millard P.H, 1994). As such, short stay patients who do not stay overnight are considered as staying for one day instead of zero days, thus ensuring that their workload is also taken into consideration.

The Surgical dataset is split into two subsets, where the training subset consists of 5148 randomly extracted spells and the testing subset is made up of the remaining 2575 spells. To ensure that the results are statistically robust, we perform the analysis on ten randomly extracted training and testing datasets.

Fitting GMM to LOS data

Once the data was split, various GMMs with components ranging between two and six were fitted to the Surgical LOS training data and the parameters were estimated using the EM algorithm. As the number of components of the GMM were increased, the diversity of the patient population was better reflected, Table 1. In addition, as the number of components was increased, the mean and variance of the last component also increased substantially.

When fitting the GMM's, components with a higher mean LOS contained only a small proportion of patient records, as shown by the mixing coefficient, e.g. 1.9% and 0.3% for the last component of the model with five components and six components, respectively. The six-component GMM seems to be overfitting the LOS data, this can be seen by observing that the sixth component represents less than one percent of patient spells.

Table 1. Estimated parameters for the GMMs fitted to the Surgical dataset by using the EM algorithm.

No of Components	Mean (days), standard deviation (days), mixing coefficient (%), per component					
	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}
2	3.8, 2.1, 76.7	16.9, 14.8, 23.4				
3	2.3, 0.5, 39.2	6.4, 3.1, 48.7	22.3, 18.1, 12			
4	2.2, 0.5, 38.3	5.4, 2.2, 39.4	13.4, 6.0, 18.8	39.2, 26.0, 3.5		
5	2.2, 0.5, 37.4	4.4, 1.5, 27.9	8.6, 3.1, 22.8	18.8, 7.5, 9.9	49.1, 30.2, 1.9	
6	2.2, 0.5, 37	4.20, 1.4, 25.9	8.0, 2.7, 23	16.6, 6.4, 11.1	37.5, 15.4, 2.7	95.5, 51.4, 0.3

Model selection
Using the MDL criterion, the overall description length was computed for each of the five GMMs. The value of the MDL criterion for each GMM shows that as the number of components is increased, a better representation of the LOS of patients is achieved. The MDL criterion suggests the four component model is optimal. In addition, based on the percentile analysis, the four-component GMMs provides a reasonable approximation of the LOS observations. This reaffirms that the four component model is representative of the LOS of surgical patients using the minimum number of components. Moreover, if we only consider percentile values up to 99.5, the χ^2 goodness-of-fit test shows no significant difference between the four-component GMM and the actual Surgical LOS data. As such we selected the GMM with four components.

Interpretation of model parameters
The four component GMM approximates the LOS distribution and suggests that there are four dominant patterns for the duration of stay of surgical patients, (the fourth group cannot be seen because of the small probability associated with it). The first group consists of approximately 38.3% of patients, with a mean LOS of 2.2 days. The second group consists of approximately 39.4% of patients, with a mean LOS 5.4 days whilst the third group consists of 18.8% of patients with a mean LOS of 13.4 days. Finally, the fourth group consists of 3.5% of patients with a mean LOS of 39.2 days.

The four surgical patient groups can be described as follows:

- The first group describes short stay patients who stay between a few hours to a couple of days.
- The second group describes those patients that are more complex than the short stay patients, who stay from a few days up to a week.
- The third group represents the patients who need more attention, staying an average of two weeks.
- The fourth group represents long stay patients, staying more than a month.

The variability within each component increases as the mean LOS within each component also increases. In other words, patients who belong in the first few groups have less variability in their likely LOS compared with longer stay patients. Based on the variation, it would therefore be easier to determine the LOS of shorter stay patients and harder to determine the LOS of longer stay patients. *Deriving the LOS classification scheme*
Based on the parameters of the four-component GMM, we use Bayes theorem, to determine the likelihood of a patient who has stayed for x days, belonging to a particular group. Using the probabilities defined for each LOS x, we assign LOS observations to the most probable group and derive a four class LOS classification scheme, namely 1-3 days, 4-9 days, 10-28 days, and 29+ days. The percentage of patients belonging to each class differs from the groups defined in the GMM. This is because the probabilistic grouping has been partitioned into non-overlapping classes.

The first LOS class within the classification scheme comprises of all patients staying between one and three days and represents 45.9% of the population. The second LOS class corresponds to patients staying between four and nine days and represents 35.1% of the population. The third class corresponds to patients staying between ten and 28 days representing 16.3% of the population, whilst the fourth class consists of any patients staying 29 days or more, representing 2.7% of the population.

Once the four-class LOS classification scheme was derived, it was then used to supervise the C4.5 algorithm in developing a surgical patient classification model. The independent variables considered within the model were limited to those available within the Surgical dataset. These included the patient's date of hospital admission, their gender, whether the patient was treated in a public or private hospital, whether the patient was admitted as an emergency case, as well as their diagnosis coded using MDC.

Classification prediction accuracy

The overall training accuracy, testing and class accuracy are detailed within Table 2. **Table 2**: Overall and class accuracy of decision trees using C4.5 (averaged over ten randomly derived trees, where standard deviation of results are shown in brackets)

Overall Prediction Accuracy		Class Accuracy			
Training (%)	Testing (%)	1-3 days	4-9 days	10-28 days	29+ days
56.0	52.1	77.4	42.9	6.4	20.4

Sensitivity analysis

In order to improve the prediction accuracy of 52.1% (Table 2), we performed a sensitivity analysis of the LOS classification scheme by modifying the interval ranges defined within the LOS classification system.

For each of the ten trees derived, we analysed the corresponding confusion matrices. The analysis indicated that although the confusion matrices differed for each tree model (depending on the data used to build and test the tree), the influence of the first two LOS classes was consistently observed. An example of this is provided in Figure 1, whereby the classifier was clearly influenced by the first class 1-3 days when predicting spells belonging to the second class 4-9 days. In this case, for patient spells belonging to class 4-9 days, 405 spells were correctly classified, however 482 spells were incorrectly classified as belonging to 1-3 days. This clearly illustrates the influence of the first class on the classifier, which could be because 45.9% of spells belonged to this class, as opposed to the 35.1% of spells belonging to the second class.

In addition, the classifier was also influenced by the first two shorter stay classes when predicting the third and fourth LOS classes. For instance, the majority of spells that belonged to class 29+ days were incorrectly classified as belonging to classes 1-3 days and 4-9 days.

	1-3 days	4-9 days	10-28 days	29+ days	Total number of spells
1-3 days	962	221	3	0	**1,186**
4-9 days	482	405	10	0	**897**
10-28 days	182	211	5	13	**411**
29+ days	29	33	3	15	**80**

Figure 1: Classifications predicted for testing data (2,574 spells) for a decision tree built on randomly extracted training data (5,149 spells), where the overall accuracy is 53.8% and class accuracy is 81.1%, 45.2%, 1.2%, 18.8%.

In order to remove the influence of the first and second class, as well as to increase the prediction accuracy for the longer stay spells, we merged the third and fourth LOS classes to form a three class LOS classification scheme. This step was performed to increase the number of spells within the longer stay class.

The three class scheme is defined as 1-3 days, 4-9 days, and 10+ days, where 45.9% of spells belonged to the first class, 35.1% to the second class, and 19.0% to the third class. Although this increased the average overall accuracy from 52.1% to 52.5% (standard deviation of 0.6), the class accuracy for the longer stay class remained very low. The class accuracy was 77.7% for 1-3 days, 40.4% for the 4-9 days, and 14.1% for the 10+ days class. The low accuracy for the long stay class was caused by the classifier still being influenced by the first two shorter stay classes, see confusion matrix for the three class scheme, Figure 2. As such, we also performed a sensitivity analysis to modify the intervals of the classification scheme in order to increase the class accuracy.

	1-3 days	4-9	10+ days	Total number of spells
1-3 days	996	days	16	**1,199**
4-9 days	510	221	37	**905**
10+ days	190	405	51	**470**

Figure 2: Classifications predicted for testing data (2,574 spells) for a decision tree built on randomly extracted training data (5,149 spells), where the overall accuracy is 54.6% and class accuracy is 83.1%, 39.6%, 10.9%.

Performing the sensitivity analysis resulted in a LOS classification scheme of 1-2 days, 3-6 days, and 7+ days. However, the ensuing overall accuracy of the new classification scheme had decreased to 49.4% (with a standard deviation of 0.8), with a class accuracy of 48.0%, 53.9%, and 45.2%. An exemplar confusion matrix of a decision tree classifier, based on the updated classification scheme after the sensitivity analysis is shown in Figure 3. In terms of the overall accuracy, the performance decreased as the class accuracy for the short stay class was also reduced.

	1-2 days	3-6 days	7+ days	Total number of spells
1-2 days	370	262	148	**1,199**
3-6 days	206	496	300	**905**
7+ days	102	255	436	**470**

Figure 3: Classifications predicted for testing data (2,574 spells) for a decision tree built on randomly extracted training data (5,149 spells), where the overall accuracy is 50.6%, and class accuracy is 47.4%, 49.5%, 55.1%.

4 Discussion

In this paper we illustrated how the classification model for predicting patient LOS maybe refined in order to increase the prediction accuracy for the longer stay spells. Although the prediction accuracy was increased, the accuracy for the shorter stay classes decreased.

The methodology is inevitably limited by the quality and quantity of the data available, and as such we emphasise that the decision tree algorithm can only capture the patterns that are present within the data (Mingers J, 1989b). With deterministic data, each example within the training data can always be correctly classified from the set of independent variables (Mingers J, 1989a). However, in many real world problems where there is a degree of uncertainty present in the data, it makes it very difficult in making accurate predictions.

References

1. **Abbi R, El-Darzi E, Vasilakis C & Millard P.** (2007a), Length of stay based grouping and classification methodology for modelling patient flow. Journal of Operations and Logistics (Submitted).
2. **Abbi R, El-Darzi E, Vasilakis C & Millard P.** (2008), Length of stay based grouping and classification methodology for modelling patient flow. Journal of Operations and Logistics (To be published).
3. **Abbi R, El-Darzi E, Vasilakis C & Millard P.H.** (2007b), Intelligent methods for modelling patient flow. The International Conference on Industrial Engineering and Systems Management (IESM 2007). Beijing CHINA.
4. **Akaike H.** (1973), Information theory and an extension of the maximum likelihood principle. Petrov BN, Csaki F (eds). Second International Symposium on Information Theory. Akademia Kiado, Budapest.
5. **Cochran W.** (1977), Sampling Techniques, New York, Wiley.
6. **Dempster A.P, Laird N.M & Rubin D.B.** (1977), Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39, 1-38.
7. **Han J & Kamber M.** (2006), Data Mining: Concepts and techniques, San Francisco, USA, Morgan Kaufmann.
8. **Harper P.R.** (2005), A review and comparison of classification algorithms for medical decision making. Health Policy, 71, 315-331.
9. **Lim T, Loh W & Shih Y.** (2000), A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms. Machine Learning, 40, 203-228.
10. **Marshall A.H, Mcclean S.I, Shapcott C.M, Hastie I.R & Millard P.H.** (2001), Developing a Bayesian belief network for the management of geriatric hospital care. Health Care Management Sciences, 4, 25-30.

11. **Marshall A.H, Vasilakis C & El-Darzi E.** (2005), Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions. Health Care Management Science, 8, 213-320.
12. **Mclachlan G.J & Peel D.** (2000), Finite Mixture Models, New York, John Wiley & Sons.
13. **Millard P.H.** (1994), Current Measures and their defects. In MILLARD P.H & MCCLEAN S.I (Eds.) Modelling hospital resource use: a different approach to the planning and control of health care systems. London, Royal Society of Medicine.
14. **Miller G.A.** (1956), The Magical Number Seven, Plus or Minus Two. The Psychological Review, 63, 81-97.
15. **Mingers J.** (1989a), An Empirical Comparison of Pruning Methods for Decision Tree Induction. Machine Learning, 4, 227-243.
16. **Mingers J.** (1989b), An Empirical Comparison of Selection Measures for Decision-Tree Induction. Machine Learning, 3, 319-342.
17. **Nabney I.T.** (2004), Netlab: Algorithms for pattern recognition, Great Britain, Springer-Verlag.
18. **Perner P, Zscherpel U & Jacobsen C.** (2001), A comparison between neural networks and decision trees based on data from industrial radiographic testing. Pattern Recognition Letters, 22, 47-54.
19. **Quinlan J.R.** (1993), C4.5: programs for machine learning, London, England, Morgan Kaufmann Publishers.
20. **Rissanen J.** (1978), Modelling by the shortest data description. Automatica, 14, 465-471.
21. **Schwarz G.** (1978), Estimating the Dimension of a Model. The Annals of Statistics, 6, 461-464.
22. **Titterington D.M, Smith A.F.M & Makov U.E.** (1985), Statistical Analysis of Finite Mixture Distributions, New York, John Wiley & Sons.
23. **Walter M.** (2002), Automatic Model Acquisition and Recognition of Human Gestures. Harrow School of Computer Science. London, University of Westminster.
24. **Walter M, Psarrou A & Gong S.** (2001), Data Driven Model Acquisition using Minimum Description Length. British Machine Vision Conference. Manchester, UK.