Healthcare data mining: predicting inpatient length of stay.

Peng Liu[1]
Lei Lei[1]
Junjie Yin[1]
Wei Zhang[1]
Wu Naijun[1]
Elia El-Darzi[2]

[1] School of Information Management and Engineering, Shanghai University of Finance and Economics
[2] Harrow School of Computer Science

Healthcare Data Mining: Prediction Inpatient Length of Stay

Peng Liu, Lei Lei, Junjie Yin, Wei Zhang, Wu Naijun, Elia El-Darzi

Abstract— **Data mining approaches have been widely applied in the field of healthcare. At the same time it is recognized that most healthcare datasets are full of missing values. In this paper we apply decision trees, Naive Bayesian classifiers and feature selection methods to a geriatric hospital dataset in order to predict inpatient length of stay, especially for the long stay patients.**

Index Terms*— *NBI, LOS, Healthcare data mining

I. INTRODUCTION

Data mining has a wide use in the healthcare domain in areas such as diagnoses [1] and patient management [2], [3]. One of the main concerns in the healthcare area is the measurement of flow of patients through hospitals and other health care facilities. For instance if the inpatient length of stay (LOS) can be predicted efficiently, the planning and management of hospital resources can be greatly enhanced [4]. Data mining algorithms have also been successfully applied to predict LOS, for example see [5] and [6]. Hospital LOS of inpatients is frequently used as a proxy for measuring the consumption of hospital resources and therefore it is essential to develop accurate models for the prediction of inpatients LOS.

However, not all learning systems are suitable for health care applications. Before applying the learning system to healthcare data, one should check whether the algorithm is meeting the following criteria [1]:

1) Performance. The algorithm should be able to extract important information from the valid data and the diagnosis precision of the algorithm should be as high as possible. The accuracy of the classifier is often higher than that of the doctor while using the same descriptions.
2) Handling missing data. There are always missing information for some patients. Hence, data mining approaches should be able to run on incomplete data.
3) Handling noisy data. Sometimes, there are errors or inconsistent data in the records. Therefore data mining approaches applied to healthcare dataset should always be able to handle the noise properly.
4) Transparency. The knowledge and conclusions should be clear.
5) Explanation. The system should be able to provide further elaboration whenever the health worker needs to understand the system's suggestion.
6) Training time. Collecting patient records is expensive and time consuming. Thus, a classifier that can diagnose with fewer data is preferred.

This paper introduces new algorithms based on apply decision trees, Naive Bayesian classifiers and feature selection and organized as follows: Section 2 describes the healthcare oriented data-mining techniques; Section 3performs data mining applications to Clinics dataset; finally, Section 4 presents the conclusions of this work.

II. DATA MINING ALGORITHMS

According to the characteristics of the healthcare dataset, this paper applies one of the two most successful and widely

[1] Peng Liu, Lei Lei, Junjie Yin, Wei Zhang and Wu Naijun are with School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai, P.R. China, 200433 (e-mail: liupeng@mail.shufe.edu.cn).

Elia El-Darzi is with Health Care Computing Group, School of Computer Science, University of Westminster, London, Northwick Park, HA1 3TP, UK (e-mail: eldarze@wmin.ac.uk).

1-4244-0195-X/06/$20.00 ©2006 IEEE

used classifiers: decision tree C4.5 [7] and its successor R-C4.5s [8], Naive Bayesian classifier (NBC) [9] and its successor NBCs [10]. In addition, Naive Bayesian imputation (NBI) model [11] is used for missing data handling.

2.1. C4.5 and R-C4.5s

Decision tree C4.5 uses attribute to split samples and normally lead to high classification accuracy. Lim et al (2000) compared twenty-two decision trees, nine statistical, and two neural network algorithms in terms of classification accuracy, training time and (in the case of trees) number of leaves [12]. C4.5 is one of the classifiers, which has the best combinations in terms of error rate and speed. However, C4.5 is not very good at explanation and not too robust. It tends to produce trees with much more leaves than those from other decision trees.

To overcome some of the limitations of C4.5, Liu et al [8] proposes a robust decision tree algorithm called R-C4.5s. Based on C4.5, R-C4.5s combines branches with little classification contribution and thus resulted in building more robust and smaller trees. R-C4.5s is hence a viable learning algorithm according to the six guidelines for healthcare data mining system proposed by Kononenko et al (1998).

2.2. NBC and NBCs

NBC uses probability to represent each class and tends to find the most possible class for each sample. Though the assumption of independency cannot be satisfied thoroughly, NBC always perform well in practice. If only the probability of the correct class id higher than those of the other classes, NBC can obtain the correct classification without the exact probability distribution. In the other words, NBC is robust and insensitive to missing data [10].

Further, NBC can build better models with few training data [9]. Kononenko et al (1998) showed that NBC performed best on the whole and is suitable for medical diagnose and prediction. Healthcare dataset attributes are commonly known for conditional independency and doctors usually attempted to define this kind of attributes. NBC is transparent and good at explanation. These futures satisfy the doctors' requirements.

NBC, however, cannot automatically select suitable features like decision trees hence the performance of NBC lies on the suitability of the features selection in dataset. Thus careful features selection can improve the performance and efficiency for NBC. Liu [10] proposed a modification for NBC, called NBCs. In this important features for the NBC are chosen using feature selection strategy based on decision tree structure.

2.3. NBI Models

Imputation technique is one of the widely used missing data treatment methods [11]. The basic idea of Naive Bayesian Imputation (NBI) is first to define the feature to be imputed, called 'imputation feature' and then construct the NBC using the imputation feature as the class feature. Other features in the dataset are used as the training subset. Hence the imputation problem is becoming a classification problem. Finally, the NBC is used to estimate and replace the missing data in the imputation feature. In this paper we experimented with four different NBI strategies and are listed below.

1) NBI-A. All the imputation features are imputed by NBI model.
2) NBI-P. Only the main imputation features are imputed by NBI model and the rest are handled by the internal methods of C4.5 (C4.5-BI).
3) NBI-I-A. The improved NBI model imputes all the imputation features. Strategy (1) uses feature selection strategy to improve the NBI model. That is, NBCs take the place of NBC in this strategy.
4) NBI-I-P. The improved NBI model only imputes the main imputation features and the rest are handled by the internal methods of C4.5 (C4.5-BI). Strategy (2) uses feature selection strategy to improve the NBI model. That is, NBCs take the place of NBC in this strategy.

III. MODEL EXPERIMENTS

In this section we apply the data mining algorithms cited in the previous section to a real life geriatric hospital dataset, called Clinics dataset.

3.1. Clinics Dataset

The Clinics dataset contains data from a clinical computer system that was used between 1994 and 1997 for the management of patients in a Geriatric Medicine department of a metropolitan teaching hospital in the UK [13]. It contains 4722 patient records including patient demographic details, admission reasons, discharge details, outcome and LOS.

Patient LOS has an average of 85 days and a median of 17 days. For ease of analysis, the duration of stay variable was categorized into three groups: 0-14 days, 15-60 days and 61+ days (variable LOS GROUP). The boundaries LOS groups were chosen in agreement with clinical judgment to help describe the stages of care in such a hospital department. The first short-stay group (0-14 days) roughly corresponds to patients receiving acute care, i.e. patients admitted in a critical condition and only staying in hospital for a short period of time. The second, medium-stay group (15-60 days) corresponds to patients who undergo a period of rehabilitation. The third, long-term group (61+ days) refers to the 11% of patients who stay in hospital for a long period of time.

The missing data accounts for a lot in this dataset. There are 3017 instances (63.89%) that contain missing data. There are 11 features with missing data. Three of them have missing proportion over 20% and the highest is to 42.5%.

3.2. Feature Selection

Healthcare datasets always have a large amount of features. However, they are not all relevant. For example, Clinics dataset has 115 features and after initial data pre-processing analysis 30 features were retained. Therefore for healthcare datasets, one major goal of data pre-processing phase is to compress datasets and prepare them for the posterior data mining algorithms. In this paper we use feature selection strategy to compress the clinic dataset.

In this we uses 'Feature Important Factor (FIF)' based on decision tree structure to define the importance of the feature for the data mining task. All the decision tree models use the features, which contribute most for the classification task as the nodes of the tree. The features selected by the tree are important for the classification task. On the other hand, the different levels and branches also reflect the relative important degrees of the features. Therefore, we can obtain 'Feature Important Factor' on the base of the decision tree structure. This process can be described as follows: First, the feature on the root assigns 1 for its FIF. The FIF on node is equal to its father node's FIF multiplies by a proportion between the number of instances of this node and the number of instances of its father node. Quinlan's C4.5 [7] is used in this paper to compute the Feature Important Factor and is called FIF-C4.5.

The goal of data mining application in this paper is to improve the predication and classification of impatient length of stay. Hence the performance of the feature selection is judged by the final classification accuracy. In according with the descending order of the 'Feature Important Factor', we first select first one, then first two, …, and finally first several features to construct decision tree and NBC. The classification accuracies are summarized in Table I.

From Table I we could see that the feature selection has contributed to the improvement of the accuracy of the classifier. The feature selection strategy based on C4.5 structure has high impact on the improvement of the classification accuracy of NBC. It also provides the basis for NBCs.

Decision tree algorithms always select the feature with the most classification power as the current node. Other features that are statistically correlated to the feature on the node have similar classification power and the tree will not select them as potential nodes for the below level, because they cannot provide new power for classification. Hence we can expect that the features selected by the decision tree algorithm in the process of constructing the tree to not only have good classification power, but also to be statistically independent.

From Table I, all the three classifiers perform better on subset selected by feature selection strategy compared with the original dataset. The FIF based on C4.5, compressed the dataset into only seven features without any lost in

classification accuracy.

TABLE I. OPTIMUM FEATURE SUBSET AND THE IMPROVEMENT OF ACCURACY FOR EACH CLASSIFIER FOR FEATURE SELECTION FIF-C4.5

Classifier	Classification Accuracy (%)			Attr. Num. for Optimum subset
	All Attr.	Optimum Attr.	Improvement (%)	
NBC	53.32	54.3	1.84	10
C4.5	54.24	55.1	1.59	7
R-C4.5s	54.26	55.57	2.41	7
Ave.	53.94	54.99	1.95	8

3.3. Missing Treatment

Clinics dataset contains a lot of missing data. Missing data treatment methods area applied to Clinics to improve the accuracy of the prediction models, especially for that of the long-term group.

For evaluating the performance of these data mining algorithms Prediction Accuracy of Model, Prediction Accuracy of Class and Imputation Profit are used and defined as follows:

$$\text{Prediction Accuracy of Model} = \frac{\text{Number of correct categorized instances}}{\text{Total number of instances}} \times 100\% \quad (1)$$

$$\text{Prediction Accuracy of Class} = \frac{\text{Number of correct categorized instances in the class}}{\text{Total number of instances in the class}} \times 100\% \quad (2)$$

$$\text{Imputation Profit} = \frac{\text{Prediction accuracy - prediction accuracy of C4.5 internal method}}{\text{prediction accuracy of C4.5 internal method}} \times 100\% \quad (3)$$

In this paper NBI models are applied to handle missing data in the following way. First, define the important imputation features. Second, multiply the missing proportion (*MD%*) by the 'Feature Important Factor (*FIF*)' of each feature which containing missing data and then sort the obtained data in descending order see *MD%*FIF-C4.5* in Table II. The feature whose value of *MD%*FIF-C4.5* is above the mean is then selected as the main imputation feature those features are shown in Table II. "*FIF-C4.5*" shows the feature important factor based on C4.5 decision tree structure. According to section 3.2, feature OUT is selected as the root of the tree and the *FIF-C4.5* of OUT is 4722/4722 (1.0000). Node BARTHEL contains 1046 samples and its *FIF-C4.5* is about 1046/4722 (0.2217). All the data in *MD%*FIF-C4.5* have been normalized to take values between zeros to ones.

For the main imputation features in Table II, NBI models are used to impute missing data on these features. C4.5 trees are constructed on the imputed and non-imputed datasets and the classification accuracies for both datasets were compared. In fact, building C4.5 trees on the dataset, which wasn't imputed by NBI is equivalent to applying internal method of C4.5 to handle missing data. From the experiments, we can conclude that the internal method of C4.5 performs well in missing data handling [10]. Table III and IV show the comparison between C4.5 internal method and NBI models in terms of performance.

From Table III and IV, we can see that, NBI models performed better in improving the classification accuracy, especially for the long stay group in comparison with C4.5 internal method. The strategy of performing NBI models on the main imputation features only and the rest using C4.5 internal methods (strategies NBI-P and NBI-I-P) outperformed the strategy of performing NBI models on all the imputation features (strategies NBI-A and NBI-I-A). Both strategies NBI-P and NBI-I-P improved the accuracies of model and classes, and the corresponding imputation profits are 0.7% and 7.0%, respectively. The improvement of the classification accuracy is due to the imputation of the missing data. On one hand, it proved the effect of the missing imputation on the improvement of the classification accuracy. On the other hand, it also showed the efficiency of NBI models.

From Table III and IV, we can also see that, the improved NBI has a great improvement in the performance of the missing data imputation. For both strategies of imputing all the features and only the main features the improved NBI improved the classification accuracy by 6.99% and 6.22% respectively, see Table V.

Through the use of NBI models, the classification accuracies of medium stay patients and long stay patients have obviously increased, especially for the long stay. The imputation profit reaches 90.0% (see Table IV). The number of long stay patients is relatively small. It only accounts for 15.11% of the dataset. However, those patients tend to stay in hospital for a very long time. The longest length of stay in hospital reaches 10453 dates and the average stay is 467.2 dates, which is 85.4 for the whole patients. 82.75% of the hospital resource is employed to serve the long stay patients which only accounts for 15.11% of the total. It is significant, therefore, to accurately predict the inpatient's length of stay.

TABLE II. WEIGHTED VALUE OF MISSING PROPORTION AND FIF FOR EACH REATURE WITH MISSING DATA IN CLINICS DATASET

MD%		*FIF*-C4.5		*MD%*×*FIF*-C4.5	
BARTHEL	42.5	OUT	1	BARTHEL	1
PC	38.6	BARTHEL	0.2217	OUT	0.4787
KIN	23.5	AMONTH	0.1457	LIVES	0.098
MARITAL	17.5	AYEAR	0.1269	ACONS	0.0204
LIVES	9.5	LIVES	0.0976	PC	0.0204
OUT	4.5	ACONS	0.0519	KIN	0.0183
ACONS	3.7	EPI	0.0497	AMETHOD	0.007
AMETHOD	1.5	AMETHOD	0.0456	MARITAL	0.0043
AMONTH	0.3	KIN	0.0074	AMONTH	0.0042
AYEAR	0.3	PC	0.005	AYEAR	0.0036
EPI	0	MARITAL	0.0024	EPI	0
Ave	12.9		0.1752		0.1504

TABLE III. IMPUTATION RESULTS (CLASSIFICATION ACCURACY) OF CLINICS DATASET

Imputation Model	Accuracy for C4.5 (%)	Accuracy of Class (%)		
		Short Stay	Medium Stay	Long Stay
C4.5-BI	54.24	84.3	23.9	19
NBI-A	53.64	76.8	28.6	30.1
NBI-P	54.63	78.1	29.1	31.1
NBI-I-A	57.39	77.6	35.8	36.1
NBI-I-P	58.03	79.2	35.5	35.9

TABLE IV. IMPUTATION PROFIT OF CLINICS DATASET

Imputation Model	Imputation Profit (%)	Imputation Profit of Class (%)		
		Short Stay	Medium Stay	Long Stay
NBI-A	-1.1	-8.9	19.7	58.4
NBI-P	0.7	-7.4	21.8	63.7
NBI-I-A	5.8	-7.9	49.8	90
NBI-I-P	7	-6	48.5	88.9

TABLE V. NBI AGAINST IMPROVED NBI FOR CLINICS DATASET

Imputation Model	Accuracy Improvement for C4.5 (%)	Accuracy Improvement of Class (%)		
		Short Stay	Medium Stay	Long Stay
NBI-I-A Vs. NBI-A	6.99	1.04	25.17	19.93
NBI-I-P Vs. NBI-P	6.22	1.41	21.99	15.43

IV. CONCLUSIONS

This paper applies data mining techniques to predict inpatient length of stay in a geriatric hospital department. This dataset accounts for a lot of missing data and a large number of features. Using feature selection strategy to compress this dataset proved to be very efficient for the improvement of the classification accuracy for LOS. Applying NBI models to handle the considerable amount of missing data can greatly increase the classification accuracy of predicting LOS, especially for the long stay group. Among the NBI strategies, the strategy only imputing on the main imputation features outperform the strategy imputing all the imputation features.

REFERENCES

[1] Kononenko, I., Bratko, I. and Kukar, M. Application of Machine Learning to Medical Diagnosis, In Machine Learning and Data Mining: Methods and Applications (Eds, Michalski, R. S., Bratko, I. and Kubat, M.) Wiley, New York, 1997, pp. 389-408.

[2] Harper, P. A review and comparison of classification algorithms for medical decision making. Health Policy, Vol71,2005, pp 315-331.

[3] Ceglowski, A. Churilov, L. Wassertheil, J. Knowledge discovery through Mining emergency department data. Proceedings of the 38th Hawaii International Conference on system sciences, 2005.

[4] Marshall, A. Vasilakis, C. El-Darzi, E. Length of Stay-Based Patient Flow Models: Recent Developments and Future Directions, Health Care Management Science Journal, Vol8, 2005, pp213-220.

[5] Harper, P. A Framework for Operational Modelling of Hospital Resources, Health Care Management Science, 5, 2002, pp. 165-173.

[6] Isken, M. and Rajagopalan, B. Data mining to support simulation modeling of patient flow in hospitals, Journal of Medical Systems, 26, 2002, pp. 179-197.

[7] Quinlan J.R., C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.

[8] Liu P., Yao Z., R-C4.5 Decision Tree Model and Its Applications to Clinical Dataset, Journal of Computer Research and Development, 42 (supplement B), 2005.9, pp96-101.

[9] Ratanamahatana, C.A. and Gunopulos, D., Feature Selection for the Naive Bayesian Classifier Using Decision Trees [J]. Applied Artificial Intelligence 17(5-6), 2003. 475-487.

[10] Liu P., Researches on Missing Data Treatment Methods and Improved Robust Decision Trees, PhD Thesis, Shanghai University of Finance and Economics, 2005.

[11] Liu P., El-Darzi E., Lei L., Vasilakis C., Chountas P., and Huang W., An Analysis of Missing Data Treatment Methods and their Application to Health Care Dataset. Lecture Notes in Artificial Intelligence, Volume 3584, Pages 583 - 590, 2005.

[12] Lim, T.-S., Loh, W.-Y. and Shih, Y.-S. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, Machine Learning, 40, 203-229, (2000).

[13] Marshall, A. McClean, S. Shapcott, C. Hastie, I. and Millard, P. Developing a Bayesian Belief Network for the Management of Geriatric Hospital Care, Health Care Management Science, Vol4, 2001, pp. 25-30