Hospital Pricing Estimation by Gaussian Conditional Random Fields Based Regression on Graphs

A. Polychronopoulou and Z. Obradovic
Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, PA 19122, USA
n.polychr@temple.edu and zoran.obradovic@temple.edu

***Abstract*— Accurate estimation of what a day in a hospital costs and what the hospital charges is of high interest to many parties, including health care providers, medical insurance companies, health researchers, and uninsured patients. The problem is complex, as the cost-to-charge ratio varies greatly from hospital to hospital and over time. In addition, the cost-to-charge ratio is often not reported, and in such cases group average values from similar hospitals are used. In this study we address the problem of estimating the cost-to-charge ratio at the hospital level by utilizing structured regression on a temporal graph of more than 4,000 hospitals, observed over 8 years, constructed from the National Inpatient Sample database. In the proposed approach, the cost-to-charge estimates at individual hospitals for a certain month obtained by an artificial neural network were used as unstructured components in the Gaussian Conditional Random Fields (GCRF) model. The diagnosis codes of treatments in each hospital were used to create a similarity metric that represents correlation among hospital specializations. The estimates of cost-to-charge ratio obtained using convex optimization of the GCRF parameters on the constructed graph were much better than those relying on group average based cost-to-charge estimates. In addition, cost-to-charge ratio estimates by our GCRF model outperformed regression by nonlinear artificial neural networks.**

Keywords— Conditional Random Fields; Cost to Charge Ratio

I. INTRODUCTION

Hospital costs and charges have been in the spotlight the last few years. Several researchers have undertaken the task of identifying patterns in hospital billing policies [1,2,3]. Most of the research done is focused on estimating the actual value of the expenses incurred by a hospital in providing care, since the values usually made publicly available are the hospital charges. The relation between hospital costs and charges is usually described by the Cost-to-Charge ratio (CtCR) which is defined as the ratio of the costs divided by the charges. In this study we will focus on all payers and all patients Hospital level CtCR.

The information carried by CtCRs is broadly used by Medicare, private insurance companies, hospitals and researchers and is therefore of vital importance. Payment rates for Medicare and Medicaid, are set by law that incorporates the CtCRs rather than through a negotiation process as with private insurers. Also Medicare uses CtCRs to determine outlier payments, while private insurance companies use this ratio in their negotiation process with hospitals. Finally researchers use CtCRs as a method to convert charges to costs. The values of CtCRs are often unreported and in researchers typically have to use average values of a group of hospitals as an estimate.

Cost-to-Charge Ratio reflects the billing policy, and consequently it is highly correlated with several characteristics of each hospital, such as the hospital size, ownership and location. At the same time, the Cost-to-Charge Ratio has been shown to strongly correlate with the Diagnosis Related Groups DRGs [4], so in this study we will try to estimate the CtCR by taking into account its dependencies on the specialization of the hospital, represented by the diagnosis codes most frequently treated. Our focus is the estimation of hospital level CtCRs, using yearly attributes of 2003 to 2009 data. The proposed Gaussian Conditional Random Fields (GCRF) model uses the hospital information to build an unstructured predictor and utilizes the diagnosis codes to create a similarity metric amongst the hospitals. The accuracy of our method is shown to significantly surpass the accuracy gained using the group average values typically used.

II. DATA

The data source for this study is the Nationwide Inpatient Sample (NIS) which is an archive that stores US hospital inpatient stays. It is provided by the Agency for Healthcare Research and Quality and is included in the Healthcare Cost and Utilization Project (UCUP). The NIS is the largest all-payer, uniform and multi-State inpatient care database that is publicly available in the United States [5]. The archive is designed to approximate a 20-percent stratified sample of U.S. community hospitals and all the discharges from the sampled hospitals are included in the database. Each year the NIS provides information on approximately 5 to 8 million inpatient stays from about 1,000 hospitals. The utilized portion of the database, years 2003 to 2009, contains 14,317 records of hospitals per year and over 56 million hospitalizations.

The NIS database contains clinical and resource use information, included in a typical discharge abstract. It also contains data on total charges for each hospital. The information of how much hospital services actually cost is made available to the users by an additional element of the database, the Cost-to-Charge Ratio files. Additionally a group average CtCR is provided, the values of which are frequently used as estimates of CtCRs for hospitals with missing information. In the 2003-2009 dataset, used in this study there are a total of 4057 missing CtCR values, a number that corresponds to approximately 28% of the cases.

III. Model Description

Given the NIS archive, a machine learning model can be trained to estimate the yearly CtC Ratios of each hospital. Assuming N hospitals for a given year, our problem can be stated as the estimation of the CtCR output $\mathbf{y} = (y_1, \ldots, y_N)$ given the input $\mathbf{x} = (x_1, \ldots, x_N)$, i.e. the extracted hospital attributes for which details are given in the experimental setup section. Both the hospital level attributes and the similarity that the hospitals present based on their specialization, can be taken into account with a structured model approach. In this framework our problem can be seen as a graph, were the nodes correspond to hospitals with attributes $\mathbf{x}$. The links of this graph are weighted and undirected and the values are given by the similarity measures described in a later section. In the following we first describe our unstructured predictor. Then we introduce the GCRF model and the details of the similarity measures.

A. Unstructured Predictor

The complexity of our problem requires the selection of a predictive algorithm that can capture many kinds of non linear relationships. Artificial Neural Networks (ANNs) are a good candidate since the addition of hidden nodes and layers can lead to networks that can address problems of high complexity. In our case the feed forward Multilayer Network with one hidden layer [6], was used as the unstructured predictor. The number of hidden nodes was set to a value that maximized the accuracy of the predictor. The activation function of the hidden layer was chosen to be a sigmoid, while a linear function was assigned to the output layer. Finally the learning algorithm used was Resilient Backpropagation [7]. The various extracted attributes are used as input $\mathbf{x}$ in our unstructured predictor while the output, $R(\mathbf{x})$, can be used as a standalone predictor and also incorporated in the proposed Gaussian Conditional Random Field model as an unstructured predictor.

B. Structured Learning

A structured learning approach attempts to simultaneously predict all the outputs, given all inputs and the dependencies between the outputs. In other words, while traditional, unstructured models use information contained in x_i to predict y_i, structured learning models will use the additional information stored in y_j for all j that is related to i. This prior knowledge about relationships among the outputs $\mathbf{y}$ is usually application-specific and is in our case the similarity of the hospitals, in respect to their specialization.

C. Gaussian Conditional Random Field

Introduced by [8], Conditional Random Fields (CRF) are probabilistic models for computing the probability $P(\mathbf{y}|\mathbf{x})$ of a possible output $\mathbf{y} = (y_1, \ldots, y_N)$ given the input $\mathbf{x} = (x_1, \ldots, x_N)$. In the continuous case the conditional distribution $P(\mathbf{y}|\mathbf{x})$ of the output can be represented by an equation of the form [12]:

$$P(\mathbf{y}\,|\,\mathbf{x}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp\Big(\sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) + \sum_{j\sim i} I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})\Big) \quad (1)$$

where i~j denotes the correlation of the outputs y_i and y_j, $A(\boldsymbol{\alpha}, y_i, \mathbf{x})$ and $I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})$ are real valued functions that are known in CRF literature as association and interaction potential. The larger the value of A, the more y_i is related to $\mathbf{x}$ and the larger the value of I the more y_i is related to y_j.. Z is a normalization constant, given by:

$$Z(x, \alpha, \beta) = \int_y \exp\Big(\sum_{i=1}^{N} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) + \sum_{j\sim i} I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x})\Big)\, dy \quad (2)$$

In CRF applications, A and I are in general defined as linear combinations of a set of feature functions in terms of the K and L-dimensional parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ [8]. If the feature functions are defined as quadratic functions of $\mathbf{y}$, then $P(\mathbf{y}|\mathbf{x})$ is a multivariate Gaussian distribution, which allows efficient learning and inference [9]. Assuming that we are given K unstructured predictors $R_k(\mathbf{x})$ that predict a single output y_i and L graphical models that represent the dependencies among the nodes, then the potentials can be written as:

$$\begin{aligned} A(\boldsymbol{\alpha}, y_i, \mathbf{x}) &= -\sum_{k=1}^{K} \alpha_\kappa (y_i - R_k(\mathbf{x}))^2 \\ I(\boldsymbol{\beta}, y_i, y_j, \mathbf{x}) &= -\sum_{l=1}^{L} \beta_l e_{ij}^{(l)} S_{ij}^{(l)}(\mathbf{x})(y_i - y_j)^2 \end{aligned} \quad (3)$$

where $e_{ij}^{(l)}=1$ if i and j nodes are connected in the graph G_l and $e_{ij}^{(l)}=0$ otherwise. $S_{ij}^{(l)}(\mathbf{x})$ represents the similarity between outputs y_i and y_j and in general depends on inputs $\mathbf{x}$. The larger the value of $S_{ij}^{(l)}(\mathbf{x})$ is, the more similar the outputs y_i and y_j are. Then the CRF model of (1) can be written as:

$$\begin{aligned} P(\mathbf{y}\,|\,\mathbf{x}) = \frac{1}{Z(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})} \exp\Big(&-\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_\kappa (y_i - R_k(\mathbf{x}))^2 \\ &- \sum_{j\sim i}\sum_{l=1}^{L} \beta_l e_{ij}^{(l)} S_{ij}^{(l)}(\mathbf{x})(y_i - y_j)^2\Big) \end{aligned} \quad (4)$$

This way the exponent E of the probability distribution $P(\mathbf{y}|\mathbf{x})$ is a quadratic function in terms of $\mathbf{y}$. Therefore we can determine appropriate mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ that transform $P(\mathbf{y}|\mathbf{x})$ to the Gaussian Conditional Random Fields (GCRF) [9]:

$$P(\mathbf{y}\,|\,\mathbf{x}) = \frac{1}{(2\pi 2^{N/2}\,|\,\boldsymbol{\Sigma}\,|^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\Big) \quad (5)$$

The learning task is to choose the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to maximize the conditional log-likelihood of the set of training examples:

$$(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \operatorname{argmax}_{\alpha,\beta}\Big(\sum P(\mathbf{x}\,|\,\mathbf{y})\Big) \quad (6)$$

In the GCRF case it can be shown [9] that imposing the constraint that all the elements of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are greater than 0 can ensure the feasibility of the learning task. Adopting the technique used in [10] that applies the exponential transformation on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters to guarantee that they are positive we can convert the learning task to an unconstrained optimization problem and use standard gradient descent to solve it.

The inference task is to find the outputs $\mathbf{y}$ for a given set of observations $\mathbf{x}$ and estimated parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ such that the conditional probability $P(\mathbf{y}|\mathbf{x})$ is maximized. In the case of GCRF, since the model is Gaussian, the prediction will simply be the expected value of the distribution which is equal to the mean $\boldsymbol{\mu}$. Therefore:

$$\hat{\mathbf{y}} = \text{argmax}_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x}) = \boldsymbol{\mu} \quad (7)$$

D. Similarity Measures

In structured learning the similarity matrix S quantifies the connection among the nodes of the graph, in the sense that the larger the value of S_{ij} the more similar the values of the outputs y_i and y_j. In the context of this paper the nodes of the graphical model are hospitals and the similarity measure should describe the resemblance of the specializations of two hospitals. In the following we introduce two similarity metrics.

1) Specialization Similarity: The specialization of each hospital can be represented by the distinct diagnosis codes assigned to each treated patient. This information can be extracted per hospital and year from the NIS database. Then the similarity between two hospitals can be defined as: Let N_1 and N_2 be the number of distinct diagnosis codes treated in each of the hospitals H_1 and H_2. Also let n be the number of the diagnosis codes of H_1 that have also been treated by H_2. We can then define the similarity between the two hospitals as the normalized count of the diagnosis codes that H_1 and H_2 have in common:

$$\text{similarity} = \frac{2n}{N_1 + N_2} \quad (8)$$

This definition of hospital similarity allows three major properties to be true. First the value of the similarity does not depend on the actual values of N_1 and N_2, but on the ratio of common and total diagnosis codes. Therefore hospitals with large or small values of N_i are treated by the similarity measure equally. Furthermore the value of this measure always ranges between 0 and 1 nicely representing a percentage. Finally, the measure is symmetric in the sense that hospital H_1 is as similar to hospital H_2 as H_2 is similar to H_1.

*2) Kullback–Leibler Divergence Similarity:*The similarity of two hospitals can also be defined by comparing the actual distributions of treated diagnosis codes in the two hospitals. The Kullback–Leibler Divergence is commonly used to compare two distributions as it a measure of the information lost when one of the distributions is used to approximate the other. However it does not have the property of symmetry that a distance/similarity measure should have. Therefore we first calculate a matrix that contains the values for every pair of hospitals and then we add to the matrix its transpose. An additional normalization can transform the values and restrict them between 0 and 1. This way we get a similarity metric with all the desired properties.

The existence of the distance metric properties is not a complete evaluation criterion. We have adopted the method of variogram plots for the verification of similarity metric relevance [11]. In these plots the similarity values between nodes are plotted against the variance of the output values y_i. The smaller the variance of the values y, the more similar the values are. Therefore a similarity metric that is relevant is one whose variogram depicts a trend in which the variance of the values is dropping as the similarity is increasing.

IV. Experimental Setup

Although the focus of this study is the estimation of hospital level CtCRs, there is a large part of variability in these values that originates from standard inflation of hospitalization costs. This information, although vital, is easily predictable and is not a focal point of this study. Therefore we obtain a transformed output variable **y** by subtracting from each CtC Ratio value the yearly average.

For the input variables **x**, we use several characteristics of each hospital with which the value of CtC ratio is expected to be highly correlated, such as the hospital type, size, ownership and location. Additional attributes have also been included in the model: the hospital teaching status and multi-hospital system membership. Furthermore, CtCR is expected to be correlated with attributes that describe the hospitalization trends of each hospital, such as the number of hospitalizations, the number of major operating room procedures, the average total charges or the percentage of the various payer groups. Also included in the model are the area wage index, average number of distinct procedures in each hospitalization, average number of patient's chronic diseases, average length of stay, average income quartile, and average number of external causes of injury diagnosis codes per hospitalization. Finally, since a large variation of CtC ratio with hospital states was observed, we separated the states into five groups based on their average CtC ratio values.

These attributes are used as input **x** in our unstructured predictor, a feed forward artificial neural network with one hidden layer. The number of nodes in the hidden layer was chosen to be 4. For the selection of this free parameter, we trained a neural network on the data of 2003 and we tested it on 2004 data. We repeated this process for several values of the number of hidden nodes and the architecture that returned the best results in terms of R^2 and RMSE was chosen.

Finally, we studied the corresponding variogram plots and noticed that the simple specialization similarity is closer to the ideal similarity, as this variogram has a clear and stable monotonic fall-off. However the KL-similarity will also be evaluated in our experiments. Furthermore we set a threshold value of 0.7, below which the similarity is set to zero following [11] and also [9] that showed that sparse similarity measures outperform dense metrics. This value was chosen since both similarity measures have started their monotonic decrease at this point. Notice however, that small alterations in this threshold value do not change our experimental results.

V. Results

Utilizing the unstructured predictor and similarity measures described in the previous section we run a series of experiments. Using the attributes extracted from the NIS database for years 2004 to 2009, we train the unstructured predictor in year N-1, for N=2005,...2009. Then we use this model to estimate the CtCR of year N. Combining these results with the similarity measures we have constructed for year N we can get a final estimate of the CtCR. The results, for various values of N are shown in TABLE I. For completeness we are reporting the accuracy of the unstructured predictor as a standalone model as well as the accuracy of a model that takes

TABLE I. GCRF ACCURACY RESULTS FOR VARIOUS VALUES OF YEAR OF ESTIMATION N.

	Group Average CtCR		**Neural Networks**		**GCRF**					
					Specialization Similarity		***KL Divergence Similarity***		***Both***	
	R^2	***RMSE***	R^2	***RMSE***	R^2	***RMSE***	R^2	***RMSE***	R^2	***RMSE***
2005	0.583	0.135	0.652	0.124	**0.691**	**0.117**	0.680	0.119	0.691	0.117
2006	0.605	0.142	0.663	0.131	**0.681**	**0.128**	0.676	0.129	0.678	0.128
2007	0.609	0.140	0.739	0.114	**0.757**	**0.110**	0.753	0.111	0.757	0.110
2008	0.571	0.196	0.681	0.169	**0.686**	**0.168**	0.663	0.174	0.681	0.169
2009	0.590	0.199	0.616	0.193	**0.674**	**0.180**	0.631	0.189	0.674	0.178

into account both of the similarity measures. Finally we compare the results of our experiments with a predictor that uses the estimated value for the CtCRs as provided by the NIS database, the group average CtCR. Notice that in order to transform the reported group average into a CtCR predictor we applied the same yearly adjustment as we did in the CtCR values, by subtracting from each value the yearly average.

The results reported here, clearly indicate that the method applied in this study provides a much more accurate estimate of the CtCRs than the values of the group average, as evaluated by the NIS. In addition the structured GCRF predictor clearly outperforms the unstructured NN predictor even with the KL – Divergence similarity measure. This is a strong indication that the similarity in the specialization of the hospitals carries valuable information that should not be ignored.

For further evaluation of our model we compare the resulted accuracy of GCRF, with the simple specialization similarity, for two distinct categories of hospitals: hospitals that existed in the previous year of NIS data and were therefore used by our unstructured predictor during training phase, and hospitals that did not exist. The results are reported in TABLE II. For comparison we also report the accuracy of the NIS reported group average for the hospitals that did not exist in the previous year. It is clear that our method outperforms the commonly used one in a systematic way.

VI. DISCUSSION AND FUTURE WORK

In this study we have proposed a method for the estimation of yearly, hospital level CtC Ratio that outperforms the commonly used method of the group average. We have also identified two similarity measures that are able to offer significant improvement in the resulting accuracy. However, the information carried by the two similarity measures is not always fully utilized. We could have seen higher improvements if the model was able to identify the 'certainty' of each similarity for each node. This would be an interesting future topic with broad applications.

TABLE II. GCRF ACCURACY RESULTS FOR HOSPITALS THAT HAVE HAVE BEEN SEEN IN THE PREVIOUS YEAR AND HOSPITALS THAT HAVE NOT

	Seen Hospitals/ All Hospitals	**Group Average CtCR**		**GCRF**		**GCRF**	
		Unseen Hospitals		***Unseen Hospitals***		***Seen Hospitals***	
		R^2	***RMSE***	R^2	***RMSE***	R^2	***RMSE***
2005	82/303	0.588	0.134	0.648	0.124	0.803	0.092
2006	84/407	0.625	0.141	0.660	0.134	0.756	0.102
2007	117/422	0.709	0.118	0.733	0.114	0.801	0.101
2008	123/436	0.658	0.187	0.662	0.187	0.792	0.109
2009	107/381	0.571	0.209	0.646	0.190	0.760	0.142

[a.] The number of hospitals reported here is after removal of undefined and unreported values

ACKNOWLEDGMENT

We acknowledge partial financial support from grant #FA9550-12-1-0406 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Project Agency (DARPA). Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality provided data used in this study.

REFERENCES

[1] Wennberg, John E., Klim McPherson, and Philip Caper. "Will payment based on diagnosis-related groups control hospital costs?." The New England journal of medicine 311.5 (1984): 295-300.

[2] Grannemann, Thomas W., Randall S. Brown, and Mark V. Pauly. "Estimating hospital costs: a multiple-output analysis." Journal of health economics 5.2 (1986): 107-127.

[3] Chen, Lena M., et al. "Hospital cost of care, quality of care, and readmission rates: penny wise and pound foolish?." Archives of internal medicine 170.4 (2010): 340-346.

[4] Shwartz M, Young DW, Siegrist R. The ratio of costs to charges: how good a basis for estimating costs? Inquiry 1995;32:476–481

[5] HCUP Nationwide Inpatient Sample (NIS). Healthcare Cost and Utilization Project (HCUP). 2003-2009. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/

[6] Hagan, Martin T., Howard B. Demuth, and Mark H. Beale. Neural network design. Vol. 1. Boston: Pws, 1996.

[7] Riedmiller, Martin, and Heinrich Braun. "A direct adaptive method for faster backpropagation learning: The RPROP algorithm." *Neural Networks, 1993., IEEE International Conference on*. IEEE, 1993.

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *Proceedings of the 18th International Conference on Machine Learning*, 2001, vol. 18, pp. 282-289.

[9] V. Radosavljevic, S. Vucetic, and Z. Obradovic, "Continuous Conditional Random Fields for Regression in Remote Sensing," in *Proceedings of European Conf. on Artificial Intelligence (ECAI)*, 2010.

[10] T. Qin, T.-Y. Liu, X.-D. Zhang, D.-S. Wang, and H. Li, "Global Ranking Using Continuous Conditional Random Fields," in *Proceedings of NIPS'08*, 2008, vol. 21, pp. 1281-1288.

[11] Uversky, A., Ramljak, D., Radosavljević, V., Ristovski, K., and Obradović, Z. Which links should I use?: a variogram-based selection of relationship measures for prediction of node attributes in temporal multigraphs. Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 676-683). ACM 2013.

[12] Sutton, Charles, and Andrew McCallum. "An introduction to conditional random fields for relational learning." Introduction to statistical relational learning (2006): 93-128.