Digital Health: Changing the Way Healthcare is Conceptualised and Delivered
E. Cummings et al. (Eds.)

doi:10.3233/SHTI190764

Early Identification of Undesirable Outcomes for Transport Accident Injured Patients Using Semi-Supervised Clustering

Hadi A. KHORSHIDI[a,1], Gholamreza HAFFARI[b], Uwe AICKELIN[a], Behrooz HASSANI-MAHMOOEI[c]
[a]School of Computing & Information Systems, The University of Melbourne, Australia
[b]Faculty of Information Technology, Monash University, Australia
[c]Faculty of Medicine, Nursing and Health Sciences, Monash University, Australia

Abstract. Identifying those patient groups, who have unwanted outcomes, in the early stages is crucial to providing the most appropriate level of care. In this study, we intend to find distinctive patterns in health service use (HSU) of transport accident injured patients within the first week post-injury. Aiming those patterns that are associated with the outcome of interest. To recognize these patterns, we propose a multi-objective optimization model that minimizes the k-medians cost function and regression error simultaneously. Thus, we use a semi-supervised clustering approach to identify patient groups based on HSU patterns and their association with total cost. To solve the optimization problem, we introduce an evolutionary algorithm using stochastic gradient descent and Pareto optimal solutions. As a result, we find the best optimal clusters by minimizing both objective functions. The results show that the proposed semi-supervised approach identifies distinct groups of HSUs and contributes to predict total cost. Also, the experiments prove the performance of the multi-objective approach in comparison with single-objective approaches.

Keywords. Injury outcomes, Semi-supervised clustering, Health service patterns, Injured patients, Multi-objective optimization.

Introduction

Injury in transport accidents is an important concern in the public health area which may lead to long-term healthcare issues such as disability and morbidity, or even mortality. An appropriate distinction between transport accident injured patients in the early stages is the key to provide the most appropriate level of care [1]. The aim of this study is to allocate injured patients into distinctive groups based on minimal historical information in such a way that groups are meaningfully related to the outcome of interest. The information we harness is the health service use (HSU) of each patient. The HSU is an important source of information toward the overall recovery process [2]. The outcome of interest in this study is the total cost of recovery. We use the HSU information of patients in the first week post-injury. Thus, this study helps health service providers to identify the patient groups with a risk of high-cost medication after the first week.

[1] Corresponding Author: Hadi A. Khorshidi, The University of Melbourne, Parkville, VIC 3010, Australia; E-mail: hadi.khorshidi@unimelb.edu.au

The data we use in this study is the recorded compensation payments of injured patients in the state of Victoria, Australia. This data is recorded by the Transport Accident Commission (TAC) and stored in an administrative database called the Compensation Research Database (CRD) [3]. After data pre-processing and attribute transformation, we establish a dataset where each row represents a compensation claim, each column is a health service and the values show how much money was spent on each health service for a claimant within the first week post-injury. The dataset contains 20,692 claims and 19 health services. These services are categorized in eight standard Medicare categories [4] as ambulance services, pharmacy, hospital services, long-term care, medical services, paramedical services, vocational and educational and other services include prosthetics and legislated post hospital support.

To reach the aim of this study, we build a semi-supervised clustering method that finds the groups of HSUs which are optimally separated from each other and have a significant association with the total cost. In a semi-supervised clustering, both unsupervised and supervised information are used [5]. Therefore, we propose a multi-objective optimization model to find the optimal clusters that minimize both the clustering cost function and the prediction error simultaneously. The research hypothesis is that a multi-objective model works better than single-objective models to find patterns. We use stochastic gradient descent and Pareto optimal solutions to find the clusters.

We implement the proposed semi-supervised method on the dataset of transport accident injured patients' claims. The implementation finds the optimal groups of HSU in the first week post-injury that are distinctive and informative about the future of patients. Also, we examine the performance of the multi-objective approach in comparison with single-objective approaches.

1. Methods

1.1. Developing the Objective Functions

One of the objective functions, which we use in our optimization framework, is the cost function for the k-medians clustering method. The k-medians method is a more robust clustering method in comparison with k-means which is sensitive to outliers [6]. For k-medians, we minimize the function formulated in (1).

$$f(X) = \sum_{r=1}^{k} \mathbb{E}[I_r(Z;X)\|Z - X^r\|_1] \tag{1}$$

where $Z \in \mathbb{R}^d$, $X = (X^1, \dots, X^k)'$ and X^r is the center of rth cluster, $\|.\|_1$ is the norm 1 and I_r is the indicator as defined by (2).

$$I_r(z;X) = \prod_{j=1}^{k} \mathbb{1}_{\{\|z-X^r\|_1 \leq \|z-X^j\|_1\}} \tag{2}$$

which means that when X^r is the nearest center to z. Our second objective function aims to examine the association between the resultant clusters and the outcome of interest. It estimates the prediction error of a multivariate linear regression model which is created by clusters (as predictors) using cross-validation (CV). In the regression model, Y is the vector of the outcomes and I is the matrix of predictors. This matrix is created by cluster centers where I_{ik} is one if ith record is allocated to kth cluster; otherwise it is zero. So, the second objective function minimizes the *m*-fold CV estimation of root-mean-square error (RMSE) as (3).

$$g(X) = \frac{1}{M}\sum_{m=1}^{M}\sqrt{\frac{\sum_{i\in F_m}(y_i-\hat{y}_i)^2}{n_m}} \tag{3}$$

where *M* is the total number of folds and F_m is the *m*th fold in the dataset, n_m is the number of records in F_m, y_i denotes the actual values of outcome and $\hat{y}_i$ represents the predicted values which are dependent on *X* and calculated by (4).

$$\hat{Y} = [1, I]_{n\times(k+1)}\beta_{(k+1)\times 1} \quad \forall i \notin F_m \tag{4}$$

which β is estimated via $([1, I]'[1, I])^{-1}[1, I]'Y$. As a result, our proposed semi-supervised clustering method is going to solve the optimization problem as (5) where *X* is the set of cluster centers.

$$\begin{cases} Min\, f(X) \\ Min\, g(X) \end{cases} \tag{5}$$

1.2. Finding the Optimal Solutions

To solve the optimization problem, the set of cluster centers is iteratively updated. We use the stochastic gradient *k*-medians algorithm that is presented in [7]. So, the cluster centers are updated using (6).

$$X_{h+1}^{r} = X_h^r - a_h^r I_r(z; X)\frac{x_h^r - z}{\|x_h^r - z\|_1} \tag{6}$$

where X_h^r is the center of *r*th cluster in *h*th iteration, a_h^r is the learning rate ($a_h^r = \frac{c_\gamma}{(1+c_\alpha n_r)^\alpha}$), n_r is the number of data points in the *r*th cluster and α, c_γ and c_α are parameters with constant values. The resulting clusters depend on the choice of the initial cluster centers [7, 8]. To overcome this drawback, we use an evolutionary computation approach. We initialize the iteration process with a pool of the sets of cluster centers. The number of clusters for each set is different. As a result, the optimization model can find the best clusters in terms of both clustering cost function and regression CV error and determines the best number of clusters.

For each iteration, we calculate both objective function values for every set. As there are two objective functions, we use Pareto optimal (non-dominated) solutions to find the best sets. These solutions have not been dominated by any other solution [9, 10]. In our case, Pareto optimal solutions are the sets where there is no other set having lower values in both objective functions. We determine the Pareto optimal sets for each iteration. We keep them in the next pool, but we update the other sets using (6). The iteration process continues until to reach the prespecified maximum number of iterations (tau_{max}). In the last iteration, the best set of cluster centers is selected among the Pareto optimal solutions. To reach this goal, we first normalize the objective function values for the Pareto optimal sets using (7) [11].

$$NormPOS = \frac{POS_{po} - \overline{POS_o}}{sd(POS_o)} | p = 1,2, \dots, n_p; o = 1,2 \tag{7}$$

Then, we select the most similar Pareto optimal solution to the positive ideal solution which is formulated as (8). We use the cosine similarity measure to the most similar set as (9) [12, 13].

$$PI = \left(\min_{o} NormPOS_{po} \,|p = 1,2,\dots,n_p; o = 1,2\right) = (PI_1, PI_2) \qquad (8)$$

$$Similarity(NormPOS_p, PI) = \cos(\theta) = \frac{\boldsymbol{NormPOS_p.PI}}{\|NormPOS_p\|_2 . \|PI\|_2} \qquad (9)$$

where PI is the positive ideal vector, POS_{po} refers to the oth objective function value for the pth Pareto optimal solution, n_p is the number of Pareto optimal solutions in the last iteration, $\overline{POS_o}$ and $sd(POS_o)$ are the mean and standard deviation values of the oth objective function for Pareto optimal solutions, $\|.\|_2$ is the norm 2 and $\boldsymbol{NormPOS_p.PI}$ is the inner product of two vectors.

2. Results

We implement the proposed semi-supervised clustering on our data set for the first week post-injury HSUs of the patients who are injured at transport accidents. The total cost to recovery is considered as the outcome of interest. Our aim is to find the separate groups of patients which are associated with different total costs. The patterns of HSU in each group are helpful to categorize the patients (in terms of total cost) in early stages (the first week). As a result, the group of patients who may lead to costly medication can be identified for special consideration.

For this implementation, we generate 100 sets of cluster centers to enter the pool (n_{pool}=100). The number of cluster centers varies from 2 to 11 (K=11). The iteration process continues for 500 times (tau_{max}=500). For the learning rate, we set parameters as $\alpha = {}^{3}/_{4}$, $c_\alpha = 1$ (as recommended in [7]), and $c_\gamma = 2000$ (based on experiments).

There are 36 Pareto optimal solutions at the end of the algorithm. We find the most similar one compared to the positive ideal solution. The final optimal set contains three cluster centers. Figure 1 shows different values of cluster cost function and regression CV error through the iterations for the optimal set.

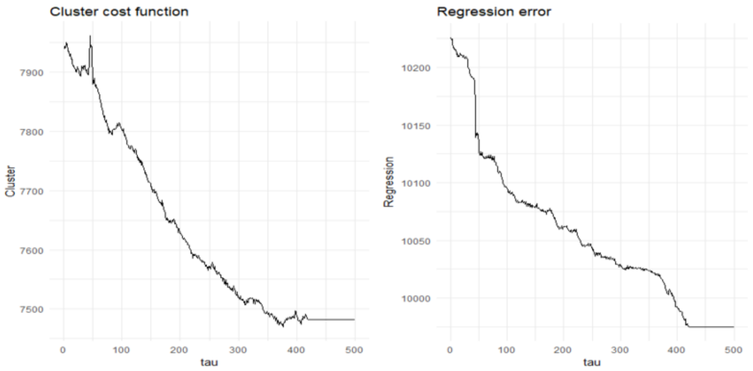

Figure 1. Cluster cost function (left) and regression CV error (right)

2.1. Validation

To examine the performance of the proposed multi-objective model, we compare the resultant clusters with the resulting clusters from single-objective models. Therefore, we find the best sets of cluster centers when we just have one of the objective functions. We use two validity indices, average silhouette width (ASW) [10] and CV error for prediction of the logarithm of total cost using regression, for the comparison. The ASW value lies between -1 and 1 and the higher value demonstrates the better partitioning results. Table 1 presents the comparison results through the index values of the optimal solution, the average and standard deviation (SD) of the last pool. The comparison results show that the multi-objective model provides significantly better portioning results. The optimal solution of the model with the objective function of regression error has the lowest CV error. However, the ASW value for this set of cluster centers is very low. In addition, the average value of CV for the multi-objective model is lower than the single-objective ones.

Table 1. Comparison results

Approach	**Multi-objective**		***k*-Medians**		**Regression**	
Index	ASW	CV	ASW	CV	ASW	CV
Optimal Solution	0.505	1.239	0.109	1.527	-0.019	1.199
Average	0.109	1.247	0.019	1.359	-0.011	1.266
SD	0.246	0.095	0.267	0.091	0.236	0.106

2.2. Interpretation

The distribution of patients in clusters is presented in Table 2. The table also brings the information of average total cost for the patients in each cluster. We can see that different clusters lead to different levels of total cost.

Table 2. The size and average total cost of clusters

Cluster	1	2	3
Number of patients	6,396	12,206	2,090
Average total cost (A$)	4,682	1,557	25,944

Figure 2 visualizes use of health services by patients in each cluster in their first week post-injury. It shows that medical (pathology, imaging, attendance and therapeutic) and hospital services shape the clusters.

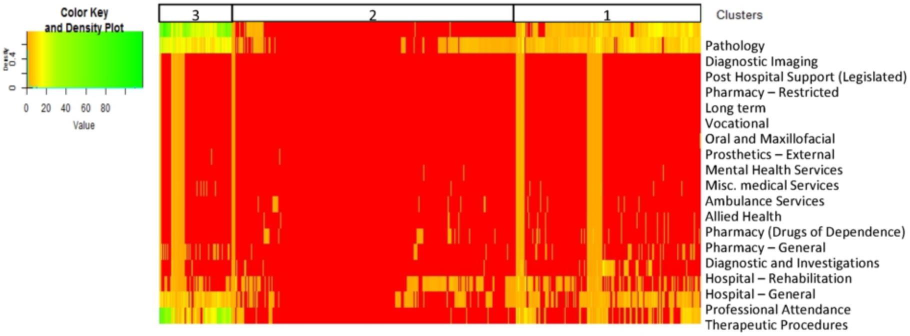

Figure 2. Heatmap for HSU in the first week post-injury

3. Conclusion

In this paper, we propose a semi-supervised clustering approach to group transport accident injured patients based on both similar health service patterns in the first week post-injury and total cost of medication. This grouping helps health service providers to identify groups with undesirable outcomes as early as possible. For semi-supervised clustering, we develop a multi-objective optimization model with two objective functions of clustering and regression. We use stochastic gradient descent and Pareto optimal solutions in an iterative learning algorithm to find optimal solutions. The implementation of the proposed semi-supervised method shows that the resultant clusters identify distinct groups of HSUs and contribute to predict total cost. Also, we examine the performance of the multi-objective model using ASW and CV indices. The examination proves that the research hypothesis is true, and the multi-objective approach outperforms the single-objective approaches. The results denote that using medical and hospital services in the first week post-injury is critical in grouping patients. For future work, the proposed semi-supervised clustering can be applied in other healthcare projects.

Acknowledgement

This project was funded by the Transport Accident Commission (TAC) through the Institute for Safety, Compensation and Recovery Research (ISCRR).

References

[1] L.J. Scheetz, J. Zhang, and J. Kolassa, Classification tree modeling to identify severe and moderate vehicular injuries in young and middle-aged adults, *Artificial Intelligence in Medicine* **45** (2009), 1-10.

[2] R.J. Mitchell, C.M. Cameron, and R. McClure, Patterns of health care use of injured adults: A population-based matched cohort study, *Injury* **48** (2017), 1393-1399.

[3] K.H. Prang, B. Hassani-Mahmooei, and A. Collie, Compensation Research Database: population-based injury data for surveillance, linkage and mining, *BMC Research Notes* **9** (2016), 1-11.

[4] Department of Health, *MBS online*. 2017: http://www9.health.gov.au/mbs/search.cfm?adv=1.

[5] S. Saha, A. Ekbal, and A.K. Alok. *Semi-supervised clustering using multiobjective optimization*. in *Proceedings of the 2012 12th International Conference on Hybrid Intelligent Systems, HIS 2012*. 2012.

[6] L.A. García-Escudero, et al., A review of robust clustering methods, *Advances in Data Analysis and Classification* **4** (2010), 89-109.

[7] H. Cardot, P. Cénac, and J.M. Monnez, A fast and recursive algorithm for clustering large datasets with k-medians, *Computational Statistics and Data Analysis* **56** (2012), 1434-1449.

[8] F. Khan, An initial seed selection algorithm for k-means clustering of georeferenced data to improve replicability of cluster assignments for mapping application, *Applied Soft Computing Journal* **12** (2012), 3698-3700.

[9] M. Abouei Ardakan and M.T. Rezvan, Multi-objective optimization of reliability–redundancy allocation problem with cold-standby strategy using NSGA-II, *Reliability Engineering & System Safety* **172** (2018), 225-238.

[10] A.K. Alok, S. Saha, and A. Ekbal, Semi-supervised clustering for gene-expression data in multiobjective optimization framework, *International Journal of Machine Learning and Cybernetics* **8** (2017), 421-439.

[11] G.W. Milligan and M.C. Cooper, A study of standardization of variables in cluster analysis, *Journal of Classification* **5** (1988), 181-204.

[12] G. Forestier, et al., Finding discriminative and interpretable patterns in sequences of surgical activities, *Artificial Intelligence in Medicine* **82** (2017), 11-19.

[13] S. Nikfalazar, H.A. Khorshidi, and A.Z. Hamadani, Fuzzy risk analysis by similarity-based multi-criteria approach to classify alternatives, *International Journal of Systems Assurance Engineering and Management* **7** (2016), 250-256.