Domain knowledge Based Hierarchical Feature Selection for 30-Day Hospital Readmission Prediction

Sandro Radovanovic[1], Milan Vukicevic[1,2(✉)], Ana Kovacevic[1,2], Gregor Stiglic[3], and Zoran Obradovic[2]

[1] University of Belgrade, Faculty of Organizational Sciences, Jove Ilića, 154, Belgrade, Serbia
{sandro.radovanovic,milan.vukicevic}@fon.bg.ac.rs
[2] Temple University, Philadelphia, PA, USA
{ana.kovacevic,zoran.obradovic}@temple.edu
[3] University of Maribor, Maribor, Slovenia
gregor.stiglic@um.si

Abstract. Many studies fail to provide models for 30-day hospital re-admission prediction with satisfactory performance due to high dimensionality and sparsity. Efficient feature selection techniques allow better generalization of predictive models and improved interpretability, which is a very important property for applications in health care. We propose feature selection method that exploits hierarchical domain knowledge together with data. The new method is evaluated on predicting 30-day hospital readmission for pediatric patients from California and provides evidence that a knowledge-based approach outperforms traditional methods and that the newly proposed method is competitive with state-of-the-art methods.

Keywords: Re-admission · Feature selection · Domain knowledge

1 Introduction

Hospital re-admissions, one of the major costs of hospital care, often result from preventable errors associated with discharging patients, such as hospital acquired infections, poor planning for follow up care, inadequate communication of discharge instructions, and failure to reconcile and coordinate medications. Timely identification of potential readmissions can have high impact on improvement of healthcare services for patients, by reducing the need for unnecessary interventions and hospital visits, as well as for hospitals, by reducing costs and improving hospital status. Algorithms for prediction of hospital re-admission often fail to produce well performing models because of high dimensionality of data (over 14,000 possible diagnoses in ICD-9 coding) and high level of data sparsity. Additionally, high dimensionality reduces the interpretability of predictive models. This is why it is utterly important to develop efficient feature selection methods that will lead to parsimonious predictive models: ones that will select a low number of features without loss in predictive performance. Even though there is a large number of current state-of the art feature selection techniques [1], only a few [4, 5] exploit domain knowledge represented in hierarchical features.

J.H. Holmes et al. (Eds.): AIME 2015, LNAI 9105, pp. 96–100, 2015.
DOI: 10.1007/978-3-319-19551-3_11

We address this problem by proposing a method that utilizes domain knowledge in the form of ICD-9 hierarchy together with data driven classification techniques. The effectiveness of the proposed approach on predicting readmission on pediatric data from California is evaluated. Additionally, we demonstrate synergetic effects of our method with Lasso logistic regression and show that it outperforms alternative methods.

2 GHFCS – Group Hierarchical Feature Compression and Selection Method

We propose a GHFCS method that exploits domain knowledge in the form of ICD-9 hierarchy of diseases, where one disease can be categorized at most in four levels, from concrete diagnosis (i.e. mononucleosis) to high level concept (i.e. infectious or parasitic disease). The main intuition behind GHFCS is that the most of the specific concepts (features) in hierarchy do not bring good quality information about observed phenomena (in our case, readmission risk). This intuition applies on EHR because of high dimensionality and sparsity of hierarchy (only a small number of examples have the same diagnosis on the most specific level), and this often leads to poor predictive performance of the algorithms. Based on this, GHFCS tends to identify features with high information potential on the highest levels of the hierarchy without losing predictive power. Instead of selecting highly specific diagnoses, we can aggregate those diagnoses to a category from a higher level of the hierarchy. If the higher level category is equally or more informative, it will be used instead of specific categories.

The GHFCS method is based on a bottom up greedy strategy and utilizes all ICD-9 hierarchical levels. First, the dataset is aggregated and fused on each level of hierarchy, creating an augmented feature space where every node in the hierarchy is represented as a feature. Further, greedy filter selection is applied starting from leaves of the hierarchy and comparing them with their parent node based on information theoretic measures (note that any information theoretic measure [1] can be used for assessment of information potential). If the average information potential of child nodes is lower than that of the parent node, then all of the child nodes are removed from hierarchy (only the parent stays as a higher concept). If opposite, the parent node is removed from hierarchy and all of the child nodes are connected to the upper level node (parent of their original parent). This allows preservation of high information potential of low level features and examination of their synergetic influence with features of higher levels. Thus, the greedy assumption is reduced.

In order to evaluate GHFCS we developed a benchmark method: **SHFCS** (Single Hierarchical Feature Compression and Selection). Unlike GHFCS, this strategy tends to keep many more features by comparison of information potential of each child node with its parent (single comparison). We also evaluate current state-of-the art methods with similar strategies. **GTD** (Greedy Top Down) [4] uses a greedy top-down approach and the most informative feature from each hierarchy path. This approach selects features in a vertical manner, and in contrast to GHFCS does not utilize the whole hierarchy (it is ignoring the fact that one feature can be present in more than one hierarchy path). **SHSEL** (Simple Hierarchical Selection) [5] identifies and

filters out the ranges of nodes with similar relevance in each branch of the hierarchy, (difference in information potential between a child and parent feature). Further, it selects only features that have greater information than the average of the tree path. Complete hierarchy is utilized, but in contrast to GHFCS, it restricts comparison to children and parent features (does not allow comparison between nodes from different hierarchy paths).

3 Experimental Evaluation

Data: Evaluation is performed on pediatric patient data (HCUP, [2]) from California about 30-day hospital re-admission containing 851 features on the lowest level of hierarchy. Data from January 2009 through December 2010 were used for training (46,682 examples), and 2011 data were used for testing (20,312 examples). Data had 14,000 binary valued features (diagnoses) and high class imbalance (11,884 positive and 55,810 negative cases). Detailed information about the dataset can be found in [6].

Comparison Between Knowledge Based and Traditional FS Methods: In the first experiment we compared the performance of knowledge-based against not-knowledge-based state-of-the-art feature selection techniques (Gini, Relief, ReliefF and MRMR) by means of Area Under Curve (AUC), Feature Space Compression (FSC) and Harmonic Mean (HM) of previous measures (where larger values are better). Figure 1 (left) shows that knowledge based feature selection methods, including our proposed Group method, are better in terms of AUC than methods that do not use domain knowledge.

On Figure 1 (middle) it can be seen that Gini has the lowest FSC by far and that not-knowledge-based methods do not follow the parsimony principle: by increasing FSC, AUC is drastically reduced. On the other side, knowledge based methods give parsimonious solutions which can clearly be seen by inspecting HM on Figure 1 (right). It is important to note that Gini was used as a measure of information potential in knowledge-based methods, which clearly shows the value of utilizing domain knowledge.

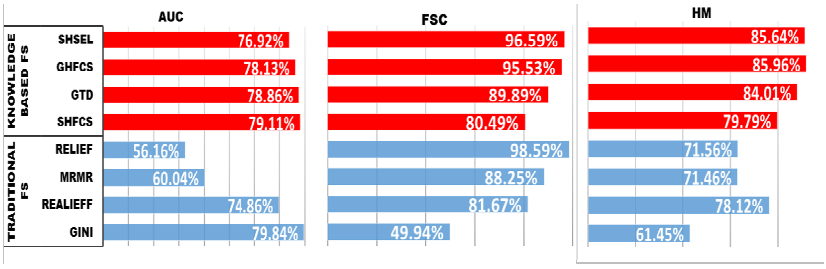

Fig. 1. Comparison of knowledge-based versus not-knowledge based FS methods

Based on results summarized at Figure 1 and for the sake of clarity of presentation in further experiments, we exclude not-knowledge-based methods from further analyses.

Comparison Between Knowledge Based Methods and Integration with Lasso: Since regularization techniques showed good performance on high dimensional sparse and imbalanced problems [3] we investigated the potential of combining Lasso regularized feature selection with knowledge based techniques. Figure 1 (left) shows the levels of AUC and FSC levels obtained from Lasso regression on subsets of features selected by knowledge based methods, as well as on the complete set of features (Complete) It can be observed that there are no significant differences in AUC. This is confirmed by significance testing between all methods (based on 100 times repeated holdout evaluation of Lasso regression on every). This result points out that GHFCS and SHSEL lead to the most interpretable solutions without loss of predictive accuracy. In order to better characterize methods, we simulated situations where AUC is more important than feature compression rate. The results are measured by HM as suggested in [5]. In our Lasso LR based experiments (summarized at Figure 2), importance of accuracy vs. interpretability (X-axis) is varied from 1 when AUC and FSC are equally important up to 5 where AUC is drastically more important. GHFCS shows the best performance in each situation (when AUC is up to 5 times more important). SHSEL showed reduced performance when AUC was more important and significantly lower performance (on 95% level) compared to GHFCS over all settings.

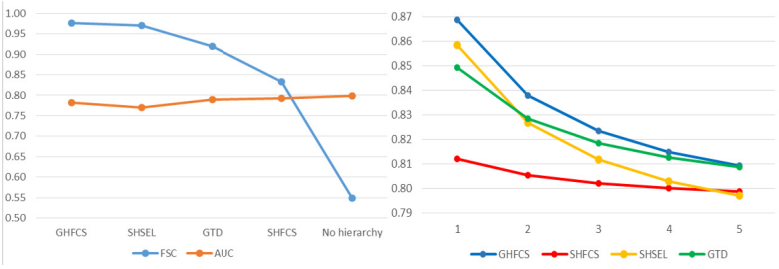

Fig. 2. Comparison between knowledge based methods

4 Conclusion and Future Research

It is shown that in contrast to traditional methods, knowledge-based feature selection preserves AUC performance and highly reduces feature space. In combination with Lasso Logistic Regression, GHFCS resulted in the most interpretable result, leaving 20 features (from the initial 851). Two categories (neoplasms and symptoms and signs) are selected as 1st level features. From the 2nd level, 14 categories are selected, mostly from the respiratory system, genitourinary system, skin tissue, sense organs diseases and injuries. On the 3rd level only food/vomit pneumonitis (ICD 5070) and exam-clinical trial (ICD V707) are selected. There were no selected features at the most specific 4th level. High level of aggregation without loss of predictive performance means that whole sub-trees of diseases are not important for predicting pediatric patient readmission. In our future work we will extend GHFCS in order to use other forms of domain knowledge (ontologies) and apply the method to different prediction tasks.

Acknowledgement. This research was supported by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR, National Science Foundation through major research instrumentation, grant number CNS-09-58854, and by SNSF Joint Research project (SCOPES), ID: IZ73Z0_152415.

References

1. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering 40(1), 16–28 (2014)
2. HCUP State Inpatient Databases (SID), Healthcare Cost and Utiliza-tion Project (HCUP)
3. Jiang, B., Ding, C., Luo, B.: Covariate-Correlated Lasso for Feature Selection. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014, Part I. LNCS, vol. 8724, pp. 595–606. Springer, Heidelberg (2014)
4. Lu, S., Ye, Y., Tsui, R., Su, H., Rexit, R., Wesaratchakit, S., Liu, X., Hwa, R.: Domain ontology-based feature reduction for high dimensional drug data and its application to 30-day heart failure readmission prediction. In: International IEEE Conference Conference on Collaborative Computing (2013)
5. Ristoski, P., Paulheim, H.: Feature selection in hierarchical feature spaces. In: Džeroski, S., Panov, P., Kocev, D., Todorovski, L. (eds.) DS 2014. LNCS(LNAI), vol. 8777, pp. 288–300. Springer, Heidelberg (2014)
6. Stiglic, G., Wang, F., Davey, A., Obradovic, Z.: Readmission Classification Using Stacked Regularized Logistic Regression Models. In: Proc. AMIA 2014 Annual Symposium, Washington, DC (November 2014)