Clustering Diagnostic Profiles of Patients

Jaakko Hollmén[1] and Panagiotis Papapetrou[2](✉)

[1] Department of Computer Science, Aalto University, Espoo, Finland
jaakko.hollmen@aalto.fi

[2] Department of Computer and Systems Sciences, Stockholm University,
Stockholm, Sweden
panagiotis@dsv.su.se

Abstract. Electronic Health Records provide a wealth of information about the care of patients and can be used for checking the conformity of planned care, computing statistics of disease prevalence, or predicting diagnoses based on observed symptoms, for instance. In this paper, we explore and analyze the recorded diagnoses of patients in a hospital database in retrospect, in order to derive profiles of diagnoses in the patient database. We develop a data representation compatible with a clustering approach and present our clustering approach to perform the exploration. We use a k-means clustering model for identifying groups in our binary vector representation of diagnoses and present appropriate model selection techniques to select the number of clusters. Furthermore, we discuss possibilities for interpretation in terms of diagnosis probabilities, in the light of external variables and with the common diagnoses occurring together.

Keywords: Medical records · Binary representations · Clustering

1 Introduction

Electronic Health Records (EHR) provide a wealth of information for retrospective analysis of patients. They can be a source for validating the conformance to planned treatment, or can be used to mimic the doctor by predicting diagnoses of patients with the use of vital signs and other symptoms. The main adoption of EHRs in healthcare research has been rapidly increasing [7,18]. In contrast to traditional data sources, including spontaneous reports [12] or social media [16], EHRs comprise disparate data types and can convey critical information which could potentially allow medical practitioners to prevent critical conditions or provide a timely intervention when necessary. A wide body of research using supervised learning approaches on EHR data exists in the literature, e.g., [5,6,11,17,19], mostly focusing on critical events such as heart failure [10] or adverse drug interactions [2,9]. On the other hand, descriptive analytics approaches focusing on frequent pattern mining or subgroup discovery have been proposed.

Published by Springer Nature Switzerland AG 2019
J. MacIntyre et al. (Eds.): AIAI 2019, IFIP AICT 559, pp. 120–126, 2019.
https://doi.org/10.1007/978-3-030-19823-7_9

More specifically, several ways of improving ADE detection have been explored [1] by combining sequential pattern mining with disproportionality analysis. In particular, the use of sequential pattern mining for finding frequent sequences of drug event prescriptions have been explored, which then form the basis for the disproportionality analysis. In other words, instead of looking for unexpected drug-diagnosis pairs, the main focus is placed on extracting unexpected pairs of drug sequences and diagnoses. Since the proposed method is better suited to handle drug interactions, it is expected to handle cases where a sequential administration of interacting drugs is responsible for a certain ADE. An empirical investigation of the method has been performed using a subset of the Stockholm EPR corpus [3].

In this paper, we focus on unsupervised learning, and more specifically on the detection of cluster structure in a medical database. Specifically, we use a large database of new-born babies treated at the neonatal intensive care unit [13] in Helsinki Children's Hospital in Finland. We use the data in retrospect to explore and investigate the diagnostic profiles of patients. For this aim, use a clustering model to group the patient database to distinct patient groups, each having a particular diagnosis signature with the recorded diagnoses. We aim at identifying interpretable diagnostic profiles by characterizing the patient profiles by the most common diagnoses in the clusters.

In the rest of the paper, we describe some backgrounds of the origins of data and present related work in Sect. 2. The methodology and the experimental part of clustering the diagnostic profiles is presented in Sect. 3. In Sect. 4, we summarize our findings and conclude our paper.

2 Patient Data and Diagnoses as Profiles

The data set under study has been recorded in the Helsinki University Hospital Neonatal Intensive Care Unit (NICU) between the years 1999 and 2013. The data set in question consists of some 2000 preterm babies born in the hospital and treated in the NICU. The treatment in the intensive care unit results in a lot of data recordings that can be analyzed in retrospect. Of particular interest in the data set has been the so called very low birth weight (VLBW) infants, which by definition are babies with a birth weight of less than 1500 g. The statistics of these data are presented in more detail in [13]. In our previous work, we have explored the possibility to predict diagnoses based on the vital signs and other symptom data, based on Gaussian process classification [14].

Contrary to our previous work [13,14], where diagnoses were estimated or predicted from vital signs and symptom based data, we consider a different approach: we treat the diagnoses of NICU patients as an individual data resource and analyze the heterogeneity and the statistical dependencies within the data and the hypothesized groups in the data. For each patient, there is a list of diagnoses given to a patient during the NICU stay. We have extracted a list of diagnosed and some other variables and cross-referenced them with an ICD-10 database retrieved from a health organization THL. For our diagnostic profile, we include only those variables which are found in a standard ICD-10 database. As a result, we get a list of 437 possible diagnoses which have occurred in this data set. We must note that this is not a standardized, comprehensive list of all diagnoses but is focused on this set of patients specifically. As the list of diagnoses for a patient may vary, we seek to represent them as a unified diagnostic profile. Therefore, we represent the diagnoses for an individual patient as a list of truth values, which can be numerically represented as 0's and 1's. Our vectorial data representation has the possible diagnoses as attributes, and the binary 0–1 values denote whether a particular patient has a diagnosis (1) or not (0). In this manner, we can represent the diagnoses as a binary 0–1 vector for a patient. This gives rise to a matrix where each row of the matrix represents the diagnostic profile of a patient.

3 Experiments: Methodology and Evaluation

Recall that our data is 0–1 data collected to a vectorial representation, where each vector describes the set of diagnoses for that patient with a collection of 0's (no diagnosis) and 1's (diagnosis) for a particular code in ICD-10. The analysis may now proceed as the analysis of multivariate 0–1 data. If we would wish to describe occurrences of diagnoses together, we could extract frequent itemsets from the 0–1 data [4], or extract frequent itemsets combined with a clustering approach [8]. Here, we are doing the first steps in exploring the data and we are content by exploring the clustering structure with a clustering approach only.

During the experiments, the goal is to use the data of the diagnostic profiles described earlier and to learn a cluster model from data. For this aim, we use the k-means cluster model, where the cluster profiles are represented in terms of prototype vectors, computer locally from the clustered data [4]. A central question regarding the clustering is to choose the number of clusters in the model: an optimal choice of the model is a trade-off between the richness of representation (many clusters) and the compactness of representation (just a few clusters). As a guiding criterion, we use the silhouette index [15] computed from the clustered data and decide on the number of clusters based on a series of silhouette indices computed on the cluster model. In order to reduce the impact of individual cluster models learned from data, we compute multiple cluster solutions form different initial values and form averages of our chosen model selection criteria. We present the statistics of these figures for solutions between 2 clusters and 20 clusters. The results of the experiment for model selection is illustrated in Fig. 1.

The choice of the number of clusters is a trade-off between the richness of representation and the compactness of the result. Whereas the compactness would make the result set very interpretable, the richness of the result would make the clustered data sets very homogeneous. In order to balance between these extremes, we resort to the silhouette index [15] and select the number of clusters to be $J = 3$. We observe declining silhouette score between 2 and 3 clusters, but an increase of clusters from 3 does not seem to affect the silhouette score.

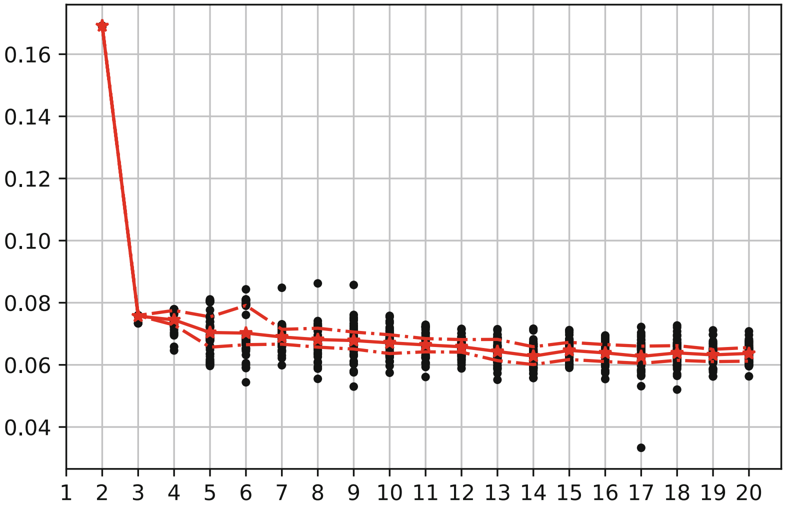

Fig. 1. The silhouette indices computed for multiple realizations of cluster models and number of clusters. We have clustered all the data in the patient matrix, with data dimension $d = 437$. We have run k-means clustering multiple times with random initializations, repeating the runs from 2 to 20 clusters 50 times. For each clustering result, we have computed the silhouette score for the solution at hand. Average of the scores calculated and plotted with a black line. Percentiles (25th and 75th) are plotted with dash dotted line. The individual scores are plotted with points.

There is, however, some variance in the score, indicated by the variance in the individual scores marked by black points as well as the bounds given by the percentiles, marked by the dash-dotted lines.

We proceed to the final clustering by fixing the number of clusters to be $J = 3$ and training a final model from data. In order to avoid degenerate results, we train a model 7 times and select the model with the median silhouette score. This is likely to avoid minima with extreme values for the silhouette index.

Since the k-means algorithm estimates the cluster centers as the averages of data and the data is either 0's or 1's, the cluster centers have a natural interpretation of being probabilities of individual diagnoses given in the cluster, and can thus be related with the risk of a diagnosis in a given group. The cluster centers for the three clusters are illustrated in Fig. 2. There are apparent similarities between the profiles and they do indeed share some characteristics in terms of diagnoses. For instance, each cluster is characterised by the diagnosis P59.0, which indicates neonatal jaundice associated with preterm delivery, and P22.9, which in turn corresponds to unspecified respiratory distress of newborn. These are diagnoses that are associated with the selection of the patient material from NICU, rather than distinguishing factors between them. Some of the diagnoses appear in only one cluster, like the H35.1 Retinopathy of maturity, which offers an avenue to explore further which factors occur together in this particular group.

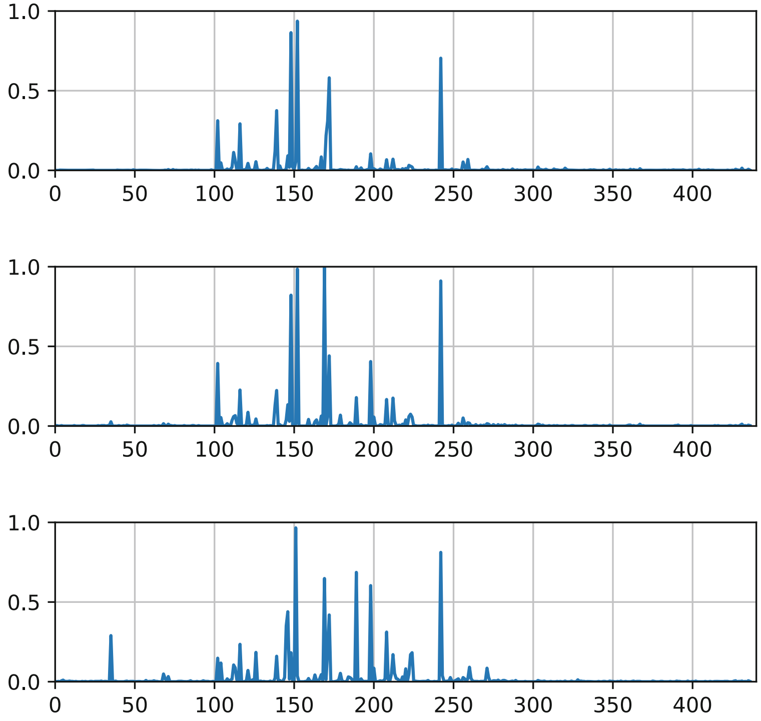

Fig. 2. The cluster centers of the identified diagnosis profiles are illustrated. Each panel describes one of the three clusters in terms of probabilities of diagnoses in the cluster.

4 Summary and Conclusions

We have analyzed a database of patients treated at the neonatal intensive care unit (NICU) at the Helsinki Children's Hospital in Finland. In particular, we focused on the set of diagnoses of patients and developed a vectorial 0–1 data representation for further analysis. The diagnostic profile for a patient is the listing of all diagnosis of a patient and can be represented as a vector of 0–1 data with all diagnoses as vector components, or attributes. Then we proceeded with a clustering approach and developed a suitable clustering model for the data through a model selection procedure. We presented the prototypes of the clusters and discussed the further possibilities of describing the data more accurately.

A clustering model can be used to yield a practical, yet nontrivial description of the patient diagnoses as such. Some of the highlighted diagnoses are generic hallmarks for the patient material in question, but some others may yield interesting information about subsets of patients. These diagnostic profiles can be used to describe further the statistical dependencies of the individual diagnoses in subsets of patients, which can yield interesting, but unexplored knowledge about the domain. In order to derive more medical relevance from the profiles, we will discuss the findings further with the medical experts, and reflect the findings with external patient data, which has not been used in clustering.

Acknowledgments. This work was partly supported by the VR-2016-03372 Swedish Research Council Starting Grant. The study was approved by the Helsinki University Central Hospital Ethics Committee, decision number 115/13/03/00/14 dated 8 April 2014. We thank Olli-Pekka Rinta-Koski for technical assistance with the data extraction.

References

1. Asker, L., Boström, H., Karlsson, I., Papapetrou, P., Zhao, J.: Mining candidates for adverse drug interactions in electronic patient records. In: Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments, PETRA 2014, Island of Rhodes, Greece, 27–30 May 2014, pp. 22:1–22:4 (2014). https://doi.org/10.1145/2674396.2674420, http://doi.acm.org/10.1145/2674396.2674420
2. Aspden, P.B.J., Wolcott J.L.R.C.: Generalized random shapelet forests. In: Committee on Identifying and Preventing Medication Errors (2007)
3. Dalianis, H., Hassel, M., Henriksson, A., Skeppstedt, M.: Stockholm EPR corpus: a clinical database used to improve health care. In: Proceedings of the Fourth Swedish Language Technology Conference (2009)
4. Hand, D., Mannila, H., Smyth, P.: Principles of Data Mining. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge (2001)
5. Harpaz, R., Haerian, K., Chase, H.S., Friedman, C.: Mining electronic health records for adverse drug effects using regression based methods. In: the 1st ACM International Health Informatics Symposium, pp. 100–107. ACM (2010)
6. Henriksson, A., Kvist, M., Dalianis, H., Duneld, M.: Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J. Biomed. Inf. **57**, 333–349 (2015)
7. Hersh, W.R.: Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance. Clin. Pharmacol. Ther. **81**, 126–128 (2007)
8. Hollmén, J., Seppänen, J.K., Mannila, H.: Mixture models and frequent sets: combining global and local methods for 0–1 data. In: Proceedings of the Third SIAM International Conference on Data Mining, pp. 289–293. Society of Industrial and Applied Mathematics (2003)
9. Ouchi, K., Lindvall, C., Chai, P.R., Boyer, E.W.: Machine learning to predict, detect, and intervene older adults vulnerable for adverse drug events in the emergency department. J. Med. Toxicol. **14**(3), 248–252 (2018). https://doi.org/10.1007/s13181-018-0667-3
10. Pakhomov, S.V., Buntrock, J., Chute, C.G.: Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. J. Biomed. Inf. **38**(2), 145–153 (2005)
11. Park, M.Y., et al.: A novel algorithm for detection of adverse drug reaction signals using a hospital electronic medical record database. Pharmacoepidemiol. Drug Saf. **20**(6), 598–607 (2011)
12. van Puijenbroek, E.P., Bate, A., Leufkens, H.G., Lindquist, M., Orre, R., Egberts, A.C.: A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions. Pharmacoepidemiol. Drug Saf. **11**(1), 3–10 (2002)
13. Rinta-Koski, O.P.: Machine learning in neonatal intensive care. Ph.D. thesis, Aalto University (2018)

14. Rinta-Koski, O.P., Sarkka, S., Hollmén, J., Leskinen, M., Andersson, S.: Gaussian process classification for prediction of in-hospital mortality among preterm infants. Neurocomputing **298**, 134–141 (2018). https://doi.org/10.1016/j.neucom.2017.12.064. http://www.sciencedirect.com/science/article/pii/S092523121830208X
15. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987). https://doi.org/10.1016/0377-0427(87)90125-7. http://www.sciencedirect.com/science/article/pii/0377042787901257
16. Sarker, A., et al.: Utilizing social media data for pharmacovigilance: a review. J. Biomed. Inf. **54**, 202–212 (2015)
17. Schuemie, M.J., et al.: Using electronic health care records for drug safety signal detection: a comparative evaluation of statistical methods. Med. Care **50**(10), 890–897 (2012)
18. Weiskopf, N.G., Hripcsak, G., Swaminathan, S., Weng, C.: Defining and measuring completeness of electronic health records for secondary use. J. Biomed. Inf. **46**(5), 830–836 (2013)
19. Zhao, J., Henriksson, A., Asker, L., Boström, H.: Predictive modeling of structured electronic health records for adverse drug event detection. BMC Med. Inform. Decis. Mak. **15**(Suppl 4), S1 (2015)