MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231083

Instance Selection Algorithms for Predictive Modelling in Telehealth Applications

Fabian WIESMÜLLER[a,b,c,1], Dieter HAYN[a,c], Florian HOFFMANN[d], Sten HANKE[b], Peter KASTNER[e], Markus FALGENHAUER[a] and Günter SCHREIER[a]
[a] *AIT Austrian Institute of Technology, Graz, Austria*
[b] *FH Joanneum, Graz, Austria*
[c] *Ludwig Boltzmann Institute for Digital Health and Prevention, Salzburg, Austria*
[d] *Versicherungsanstalt öffentlich Bediensteter, Eisenbahnen und Bergbau, Vienna, Austria*
[e] *telbiomed Medizintechnik und IT Service GmbH, Graz, Austria*
ORCiD ID: Fabian Wiesmüller https://orcid.org/0000-0001-6567-7782

Abstract. Telehealth services are becoming more and more popular, leading to an increasing amount of data to be monitored by health professionals. Machine learning can support them in managing these data. Therefore, the right machine learning algorithms need to be applied to the right data. We have implemented and validated different algorithms for selecting optimal time instances from time series data derived from a diabetes telehealth service. Intrinsic, supervised, and unsupervised instance selection algorithms were analysed. Instance selection had a huge impact on the accuracy of our random forest model for dropout prediction. The best results were achieved with a One Class Support Vector Machine, which improved the area under the receiver operating curve of the original algorithm from 69.91 to 75.88 %. We conclude that, although hardly mentioned in telehealth literature so far, instance selection has the potential to significantly improve the accuracy of machine learning algorithms.

Keywords. Instance selection, training data selection, predictive modelling, telehealth

1. Introduction

Due to current trends in healthcare, the covid-19 pandemic, and the increased use of smart technologies, an increase in the use of telemedicine and telehealth systems can be seen [1]. Telehealth plays an important role, especially in the provision of care for chronically ill patients suffering from e.g., cardiovascular or metabolic diseases such as heart failure, hypertension, chronic obstructive pulmonary disease, or diabetes. In 2010, a telehealth system called DiabMemory was developed by the AIT Austrian Institute of Technology to support the treatment of diabetes patients [2,3]. One major aim of any telehealth service for chronic disease management is keeping the patients adherent to the service and to avoid dropouts. From our experience, in many cases, it would be possible

[1] Corresponding Author: Fabian Wiesmüller, email: fabian.wiesmueller@ait.ac.at.

to motivate patients to stay adherent, if potential dropouts were detected in time. Therefore, the prediction of future dropouts would be extremely valuable.

In an ongoing project, we are currently developing a random-forest-based dropout prediction model, which is trained based on DiabMemory data. Therefore, more than 3,000 (in parts highly correlated) features were calculated for each patient for every single day that the patient was active (see chapter 2).

At the time of the analyses presented in this paper, a simplified version of this model could predict dropouts with an area under the receiver operating curve (AUROC) of approximately 0.70. Different ways of optimizing the model accuracy were identified, namely, to apply different machine learning algorithms, to calculate additional features, to optimize the feature selection and to introduce instance selection (IS) algorithms. The present paper focuses on the optimization of the IS.

Altogether, for each patient, data from each day the patient was enrolled in the telehealth program were available (up to 12 years). Approximately half of the patients were still active at the time of exporting the data for the present study. Therefore, all their data should be predicted as "negative" (non-dropout) events. For patients who dropped out, only the time right before the dropout should be predicted as a "positive" event, while all other days should be considered "negative". Therefore, the dataset was highly unbalanced with many more negative than positive events. Additionally, data of one and the same patient derived on consecutive days are highly correlated.

IS is the process of reducing the entire dataset to a drastically smaller set of highly significant instances. This subset is then used to train a machine learning model with the same, or even higher accuracy, than training with the original dataset would have achieved. In addition to balancing the dataset, previous work has proven that IS results in more stable models with a better generalization, since outliers and noise get reduced [5,6]. Additionally, less computational time is needed, since IS aims at removing redundant entries as well [5].

IS can be done based on manually implemented, "intrinsic" algorithms, e.g., by selecting data only once per week or month, or at specific time points (e.g. enrollment, follow-up visits, etc.). One of the earliest approaches of automated IS was performed with the Nearest Neighbor (NN) algorithm [7]. This approach was adapted over time and variations like the Condensed NN, the Reduced NN and the Edited NN (ENN) were developed [8]. Regarding unsupervised IS, clustering algorithms and outlier detection can lead to more balanced, less noisy datasets [9]. Another unsupervised approach would be to calculate the mutual information (MI) between instances and only include instances above a certain MI threshold [10].

Even though IS is a common problem in machine learning, very little work has been done in the setting of telehealth systems. Therefore, the present study aimed to implement, validate and compare different IS algorithms for the described, pre-existing predictive modelling software for diabetes telehealth applications.

2. Methods

2.1. Dataset

The dataset used for this work consisted of 1,240 DiabMemory patients who were active for up to 12 years. All data was pseudonymized before subsequent work. The present study was conducted in accordance with the declaration of Helsinki, and it was covered by an approval of the ethics committee of lower Austria (vote number GS1-EK-4/534-2018). Never-beginners, i.e., patients who never actively transmitted their diabetes data, were excluded from this work, resulting in 1,197 eligible patients. The following types of features were calculated per day:

- General patient data, such as age at monitoring start, gender, etc.
- Feedback data, (e.g. number of feedbacks one week prior to the respective day)
- Patient reported data, e.g., blood sugar, body weight, subjective wellbeing, insulin admission, and physical activities, including the number of data and statistical measures derived from the data (e.g., the mean value within the preceding week).

This resulted in a feature matrix of 4,200 days x 3,030 features per patient, whereas for days before the start or after the dropout of a patient, the data were set to null. By applying supervised and unsupervised feature selection methods, the matrixes were reduced to the size of 4,200 days x 344 features per patient.

2.2. Instance selection algorithms

The IS algorithms used in this work were divided into three categories, namely *intrinsic*, *supervised* and *unsupervised* IS. The algorithms were applied on each patient's feature matrix individually. Details concerning each category are provided in the following.

2.3. Intrinsic instance selection

Four intrinsic IS algorithms were implemented, that were identified to be commonly used in published telehealth papers:

- "Random selection": randomly selecting *n* instances per patient.
- "Binning": selection of instances at predefined intervals (weeks, months, etc.)
- "Dropout aligned": selection of all positive events (dropouts) on the day before the last data transmission and selection of a random instance for non-dropouts (dropout aligned approach).
- "Sampling at dropout": Sampling was done at every dropout, meaning that a cross-section of the entire dataset was selected for the days a patient dropped out.

2.4. Supervised instance selection

For the supervised IS, the previously mentioned Nearest Neighbor approach was implemented and further extended by using a regression model instead of the binary k-NN algorithm, which enabled the possibility of selecting only instances above a certain threshold of deviation from the baseline.

2.5. Unsupervised instance selection

For the unsupervised IS, a one class support vector machine (SVM) was used to separate the data into instances within and instances outside of the calculated decision boundary. The instances outside of the boundary are usually considered to be outliers [11]. However, since the available dataset was not supposed to be very noisy, different combinations of *n* instances from within and outside of the boundary were compared with approaches that only selected *n* instances from either one of the classes.

The MI was determined as the standard deviation of a seven-day moving window, only including instances with a standard deviation above a certain threshold. Finally, these approaches were combined with the ENN algorithm since the ENN works best in combination with other algorithms [12].

2.6. Training and validation

The dropout prediction algorithm was a random forest with 100 trees which was trained with a 10-fold cross-validation. The primary measure to determine the performance of the different algorithms was the area under the receiver operating curve (AUROC) which ranges from 0.5 to 1 where 1 is a perfect result and 0.5 represents a random decision.

3. Results

Table 1 depicts the results of the best performing configuration of the IS algorithms described in chapter 2. The One Class SVM achieved the overall highest AUROC of 75.78 when applied jointly with the ENN algorithm. Additionally, Table 1 shows the ratio of dropouts to non-dropouts in every feature set in the column *Event per Non-Event* and the number of instances in each feature set.

Table 1. Comparison of the performance of the instance selection algorithms with regards to the ratio of events to non-events and the number of instances.

Algorithm	Area under the receiver operating curve (in %)	Event per Non-Event	Number of Instances
Dropout Aligned	69.91	1.16	1,052
Random Selection	68.80	0.10	11,220
Binned	69.19	0.02	73,661
Sampling at Dropout	72.78	1.00	1,138
Classification	68.91	0.08	300,584
One Class SVM	75.88	0.45	21,559
Mutual Information	59.67	0.01	162,425

4. Discussion

For this work, multiple IS algorithms were tested on a large real-world dataset originating from a diabetes telehealth system called DiabMemory. After evaluating several algorithms, separated into three categories, the One Class SVM achieved the best result with an AUROC of 75.88% which outperformed the second-best approach by 3.1%. This is quite an impressive improvement as compared to the original model.

Even though the One Class SVM achieved the highest AUROC, this does not necessarily mean that it was the best IS algorithm. As can be seen in Table 1, the One Class SVM used approximately 19 times the number of instances than the sampling at dropout method. Therefore, it can be said that for use cases that rely on a small training set due to, e.g., computational power or storage restraints, methods like sampling at dropout are valuable due to the low number of instances and the, compared to other methods, high AUROC. However, if the aim of the model is to achieve the highest possible AUROC regardless of computational times, this work showed that the One Class SVM would be the method of choice.

Our results were derived from one specific dataset (DiabMemory data) based on one specific machine learning algorithm (a random forest). Future work includes testing our findings on different datasets, originating from e.g., heart failure telehealth services, to analyze whether these results are replicable. Additionally, the influence of the monitored disease and time constants of monitored data on the IS methods should be evaluated. Additionally, it should be evaluated if multiple iterations of algorithms with a random component, like e.g., the random selection, results in significantly different results in every iteration or not. Finally, we are planning to investigate the potential of IS on different learning algorithms, especially various neuronal network architectures (e.g., long-short-term memory, concurrent neuronal network, residual neuronal network).

5. Conclusions

We conclude that IS has the potential to significantly improve the accuracy of machine learning algorithms for dropout prediction, while it is currently hardly mentioned in related literature, which mainly focuses on machine learning algorithms and feature selection.

References

[1] Mahtta D, Daher M, Lee MT, Sayani S, Shishehbor M, Virani SS. Promise and perils of telehealth in the current Era. Curr Cardiol Rep. 2021 Jul;23(9):115, doi: 10.1007/s11886-021-01544-w.

[2] Von der Heidt A, Ammenwerth E, Bauer K, Fetz B, Fluckinger T, Gassner A, Grander W, Gritsch W, Haffner I, Henle-Talirz G, Hoschek S, Huter S, Kastner P, Krestan S, Kufner P, Modre-Osprian R, Noebl J, Radi M, Raffeiner C, Welte S, Wiseman A, Poelzl G. HerzMobil Tirol network: rationale for and design of a collaborative heart failure disease management program in Austria. Wien Klin Wochenschr. 2014 Nov;126(21-22):734-41, doi: 10.1007/s00508-014-0665-7.

[3] Riedl M, Kastner P, Kollmann A, Schreier G, Ludvik B. Diab-Memory: A smart phone based data service for intensified insulin therapy in patients with type 1 diabetes mellitus-a pilot study. Diabetes. 2005 Jun;54:A489.

[4] Peinado I, Villalba Mora E, Mansoa F, Sanchez A, authors Rodriguez Mañas L, Graafmans W, Abadie F, editors. Strategic Intelligence Monitor on Personal Health Systems Phase 3 (SIMPHS3). Diabmemory (Austria). Case Study Report. EUR 27171. Luxembourg (Luxembourg): Publications Office of the European Union; 2015. JRC95122I.

[5] de Haro-García A, García-Pedrajas N. Boosting instance selection algorithms'. Knowledge-Based Systems. 2014;67:342-60, doi: 10.1016/j.knosys.2014.04.021.

[6] Saha S, Sarker PS, Saud AA, Shatabda S, Hakim Newton MA. Cluster-oriented instance selection for classification problems. Inf Sci. 2022 Jul;602:143-58, doi: 10.1016/j.ins.2022.04.036.

[7] Hart P. The condensed nearest neighbor rule (corresp.). IEEE Trans Inf. 1968 May;14(3):515-6, doi: 10.1109/TIT.1968.1054155.

[8] Wilson DR, Martinez TR. Reduction techniques for instance-based learning algorithms. Mach Learn. 2000 Mar;38(3):257-86, doi: 10.1023/A:1007626913721.

[9] Zhang W, Tan X. Combining outlier detection and reconstruction error minimization for label noise reduction. In: 2019 IEEE International Conference on Big Data and Smart Computing (BigComp); 2019 Feb 27; p. 1-4. IEEE, doi: 10.1109/bigcomp.2019.8679275.

[10] Ircio J, Lojo A, Mori U, Lozano JA. Mutual information based feature subset selection in multivariate time series classification. Pattern Recognit. 2020 Dec;108:107525, doi: 10.1016/j.patcog.2020.107525.

[11] Manevitz LM, Yousef M, One-class svms for document classification, J Mach Learn Res. 2002 Mar;2;139-54.

[12] Alejo R, Sotoca JM, Valdovinos RM, Toribio P. Edited nearest neighbor rule for improving neural networks classifications. In: Zhang, L., Lu, BL., Kwok, J, editors. Advances in Neural Networks - ISNN 2010. ISNN 2010. Lecture Notes in Computer Science, 2010; Berlin, Heidelberg: Springer, doi: 10.1007/978-3-642-13278-0_39.