Liliana BYCZKOWSKA-LIPIŃSKA[1], Agnieszka WOSIAK[2]

University of Computer Sciences and Skills (1)
Institute of Information Technology, Lodz University of Technology (2)

doi:10.15199/48.2017.12.29

Instance Selection Techniques in Reduction of Data Streams Derived from Medical Devices

Abstract. *The research described in this paper concerns the reduction of streams of data derived from medical devices, i.e. ECG recordings. Experimental studies included three instance selection techniques: thresholding method, bounds checking and frequent data reduction . It was shown that application the instance selection techniques may reduce data stream by over 90% without losing anomalies or the measurements that are key values for the medical diagnosis .*

Streszczenie*. W ramach niniejszej pracy przeprowadzona została redukcja strumienia danych pozyskanych z urządzeń medycznych. Badania eksperymentalne obejmowały zastosowanie trzech technik selekcji przypadków: metody eliminacji progowej, weryfikacji zakresu oraz redukcji obiektów częstych. W pracy zostało wykazane, że zastosowanie selekcji przypadków pozwala na redukcję strumienia danych o ponad 90% bez utraty wartości kluczowych dla postawienia diagnozy medycznej. (**Redukcja strumienia danych pozyskiwanych z urządzeń diagnostyki medycznej za pomocą technik selekcji przypadków**).*

Słowa kluczowe: selekcja przypadków, strumień danych, analiza danych medycznych
Keywords: instance selection, data stream, medical data analysis

Introduction

The development of health technologies and the capabilities of diagnostic equipment make the process of medical analysis extremely challenging due to the multidimensional large datasets. Automated knowledge extraction from such huge datasets and the medical diagnosis based on the data analysis is a highly complex issue. This is primarily due to the limitations imposed by the performance of computer systems as well as the methodological problems inherent in multidimensional data analysis. It is called the curse of dimensionality [1, 2, 3, 4].

Reduction of large data sets can be performed by reduction of the number of analyzed parameters (dimensions) or by reduction of the number of analyzed cases. The dimensionality reduction can be accomplished through statistical methods, primarily Principal Component Analysis (PCA) [5, 6] or using feature selection techniques [7, 8]. The reduction of a dataset cardinality can be achieved by sampling, grouping or instance selection methods.

Data sampling involves retrieving certain elements of a collection according to the specified distribution of these elements. Data sampling techniques are used to reduce the amount of data and to approximate data characteristics by performing summarizations [9, 10].

We can distinguished two kinds of data sampling: fandom and focused [8]. Random sampling techniques assume equal probability of occurrence for each element of the input stream in the target sample. They build a good solution for large datasets, but may be less efficient for medical data sampling, where anomalies detection is crucial for diagnostics.

Medical data analysis may refer to the form of time series, i.e. measurements performed at specified time intervals (ECG signals, heart rate measurements, automated ambulatory blood pressure measurements). The process of medical data streams analysis usually focuses on the following tasks:

- anomaly detection in data stream, i.e. classification of each element into one of two categories: normal or abnormal values (binary classification); only incorrect values should be the subject of further analysis by medical experts,
- anomaly type classification, i.e. assigning each element of the stream into one of several categories: one of them represents normal values, and the other individual disease units (multi-class problem);
- determination of frequencies for abnormal values, i.e. the percentage of abnormal values in relation to the whole dataset.

Instance selection aims to reject most of the data stream elements and preserve only the most informative elements [11]. In comparison to basic sampling techniques, it depends on the task, while sampling process can be generalized [12].

The aim of the paper is to constitute an independent contribution to the relevant literature with regard to reduction of medical data streams. The research strived for finding a successful way to perform instance selection of medical data streams derived from diagnostic devices.

The experimental studies were performed to assess the process of data sampling technique with respect to the representativeness of the sample against the complete dataset. The investigation included the following stages:

- data preparation,
- the analysis of complete datasets to distinguish values above, below, and within standardized thresholds, as well as their frequencies,
- performing instance selection for each of the datasets,
- the analysis of reduced datasets to distinguish values above, below, and within standardized thresholds, as well as their frequencies,
- the validation of results – the comparison of results and the assessment of data reduction.

The remainder of this paper is organized as follows. Section 2 (Related Works) presents literature review concerning instance selection applied in the diagnosis of medical data. Next section (Data Reduction Methods) concerns the problem of dataset cardinality reduction. In Section 4 (Experimental Analysis and Results) we describe the studies that were conducted. We introduce data collected for this application and discuss the results. Finally, in Section 5 (Conclusions) we summarize our research and describe further works.

Related Works

Large number of researches, concerning big data preprocessing and handling, were discussed in the literature during the last years [12, 13]. They confirm that

these techniques may be successfully applied in medical domain. At the same time it is noticeable that very few methods are used for treatment.
The sampling techniques are successfully applied in social networks, electronic mail analysis and industrial areas of interest [12]. However their implementation for medical data analysis is still not common. Moreover, many sampling methods have not been validated on really huge datasets. Therefore, more research studies are required [13].
Further scientific investigations, including this research, regarding instance selection, may increase the chances to implement data sampling in medical practice. Consequently, in the future medical staff will be able to avoid time and space consuming analysis of numerous parameters obtained from medical studies.

Data Reduction Methods

Large amounts of data require more and more sophisticated mechanisms for their analysis. Although more data for analysis may lead to more complex diagnosis, it may use large amounts of processing time and storage [14]. Therefore a wide range of methods aiming in the complexity reduction of the data are considered. Two main approaches - feature selection and instance selection – may be distinguished (Fig. 1).

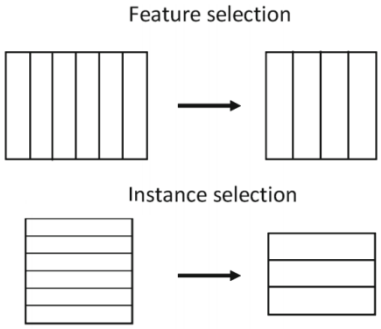

Fig. 1. Data reduction approaches.

The analysis of datasets that are numerous in instances and are characterized by lots of parameters has to cope with the problem of so-called “the curse of dimensionality” [2]. It is a serious difficulty due to the fact that the computational cost of the data processing rise and therefore it negatively impacts the actions of the most of data mining algorithms.

The process of feature selection (FS) involves the identification of the most relevant attributes from the instance’s characteristics and the removal as many as possible the irrelevant or redundant parameters” [15]. The goal of the feature selection is to build a subset of features from the original set of attributes that still reflects the relations and characteristics of the entire set of instances [16, 17]. The application of feature selection as the preprocessing technique reduces the search space defined by the features, and as a result makes the learning process faster and less memory consuming [18]. The process of feature selection might also reduce the possible risk of over-fitting in the subsequent data mining algorithms, e.g. the classification or grouping.

The process of feature selection is usually used as a preprocessing technique for other data mining algorithms, but it can also help in tasks not directly related to the data exploration. If the feature selection is applied during the data collection stage, it saves costs in time, sampling, detection and further data analysis. Moreover, models and visualizations derived from data with smaller numbers of features will be more understandable and interpretable for domain experts (e.g. for medical staff).

Another possibility for a dimensionality reduction, except for the feature selection, is application of the space transformation techniques. They are used to generate an entirely new set of features of a smaller cardinality by combining the original attributes. Many different approaches of space transformation were proposed in the scientific literature to meet different criteria and specific requirements of the domain applications. Nonetheless, the first and the most common approaches are based on the linear methods and include factor analysis and principal component analysis PCA [19, 20, 21].

Principal component analysis (PCA) is used as a standard tool in multivariate data analysis to reduce the number of space dimensions, at the same time preserve dataset's variation and dependencies [22]. The process involves the exploration of the first few data components that hold the majority of the dataset variation, instead of investigating thousands of original variables. Then, the statistical methods as well as visualization of the selected components are usually used to find similarities or differences between the samples in the considered dataset. Principal component analysis may also bring benefits also for datasets with an average or a low number of features [23].

The techniques that gain popularity in minimization of the negative impact of gathering very large datasets are instance reduction (IR) methods. They enable reduce the size of the dataset without decreasing the quality of the knowledge that can be extracted from it. Instance reduction can be used as a complementary task regarding feature selection techniques. This is due to the fact that the instance reduction decreases the quantity of data by removing instances, while the process of feature selection reduces the number of attributes (see Fig. 1).

Nowadays, the instance selection process is considered as a key stage of large datasets analysis [24]. The main difficulty for the application of the instance selection is the identification of suitable cases from a very large amount of instances. It should be performed in such a manner that there will be no information lost while carrying out the subsequent algorithms. Therefore, instance selection consists of the techniques that choose the best subset of cases that can replace the original dataset and, what is more, the new subset can fulfill the goal of a data mining

application as the original dataset [25, 26]. Moreover, a distinction between instance selection and simple data sampling should be done. Data sampling in its basic form constitutes a randomized choice of cases from the original dataset, whereas instance selection involves more complex operations to categorize cases and pick only the ones of a special value for the further analysis [27]. Additional benefits of the instance selection applied during the preprocessing stage include the dataset cleaning based on removal of noisy and redundant instances and, as a consequence to focus on the most important part of the data.

There are many algorithms that select instances taking into account their characteristics, are comprehensive, and may be applied in medical data analysis [28].

Instance-Based learning Algorithm 3 (IB3) introduced in [29] is based on accuracy and retrieval frequency measures. Removal of the cases is performed whenever the accuracy of a case is below its class frequency with a certain degree of confidence..

The Locally Weighted Forgetting (LWF) algorithm described in [30] is on k-nearest neighbors approach. The

cases with a weight below a threshold are removed. However, sometimes it may lead to overfitting [31].

Iterative Case Filtering Algorithm (ICF) proposed in [32] refers to a redundancy removal technique. It rejects the instances that have a coverage set size smaller than its reachability set.

Experimental Analysis and Results

Data for the analysis were obtained from the research server provided by National Institute of General Medical Sciences [34, 35]. They referred heart arrhythmia. The following datasets were analyzed::

- Congestive Heart Failure RR Interval Database,
- Spontaneous Ventricular Tachyarrythmia Database.

The first dataset – Congestive Heart Failure RR Interval Database – refers to beat annotation files for 29 long-term ECG recordings of subjects, aged 34 to 79, with congestive heart failure (NYHA classes I, II, and III). The recordings were digitized at 128 samples per second, and the beat annotations were obtained by automated analysis with manual review and correction. The original ECG recordings are not publically available. The interval histogram for one patient is shown in the Figure 2.

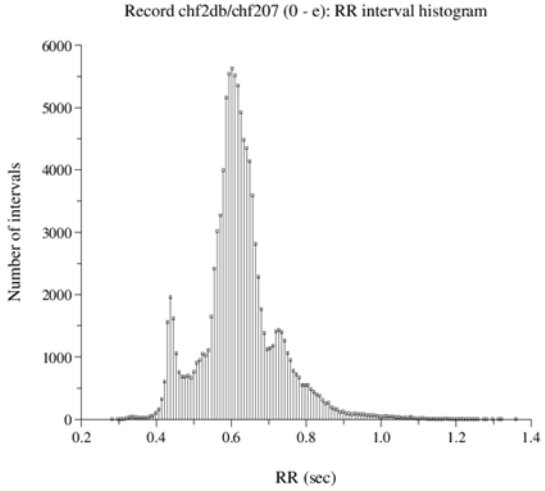

Fig. 2. The sample RR histogram for a record of Congestive Heart Failure RR Interval Database [Source: PhysioBank's Automated Teller Machine].

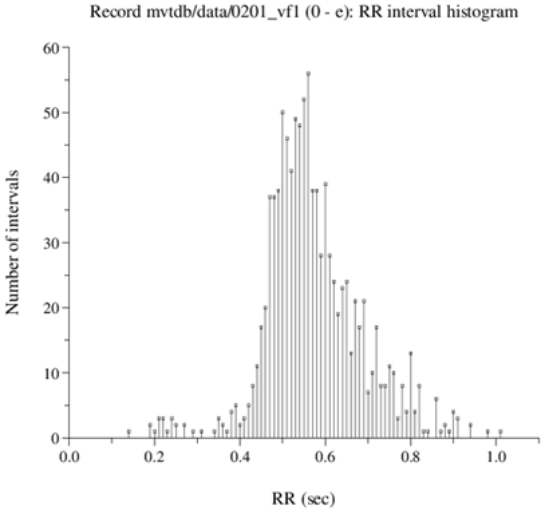

Fig. 3. The sample RR histogram for a record of Spontaneous Ventricular Tachyarrythmia Database [Source: PhysioBank's Automated Teller Machine].

The second dataset – Spontaneous Ventricular Tachyarrythmia Database – includes 135 pairs of RR interval time series, recorded by implanted cardioverter defibrillators in 78 subjects. Each series contains between 986 and 1022 RR intervals. One series of each pair includes a spontaneous episode of ventricular tachycardia (VT) or ventricular fibrillation (VF), and the other is a sample of the intrinsic rhythm. The interval characteristics for a selected record is shown in the Figure 3.

The experimental studies consisted of the following steps:

- the analysis of complete datasets to distinguish values above, below, and within standardized thresholds, as well as their frequencies,
- performing instance selection for each of the datasets - application of instance selection techniques,
- the analysis of reduced datasets to distinguish values above, below, and within standardized thresholds, as well as their frequencies,
- the validation of results – the comparison of results and the assessment of data reduction.

The following, unsupervised techniques have been considered in conjunction to obtain the best results in terms of space reduction and further data analysis time [36, 37]:

- thresholding method,
- bounds checking,
- frequent data reduction.

Tables 1. and 2. shows the numbers of instances and the gain for the proposed methodology. The first column of each table ("*All*") contains the numbers of instances for the entire datasets, the second column ("*Selected*") refers to the numbers of instances after performing instance selection actions, and the third column ("Gain") the percentage profit of the reduction.

Table 1. Numbers of samples before and after instance selection for the Congestive Heart Failure RR Interval Database.

All	Selected	Gain [%]
3 207 121	100 586	96,86 %

Table 2. Numbers of samples before and after instance selection for the Spontaneous Ventricular Tachyarrythmia Database.

All	Selected	Gain [%]
78 156	6 098	92,20 %

One can see, that for each dataset was reduced by over 90% of instances. This may lead to the conclusion that if the technique is applied, the disk space taken by data storage can be significantly reduced. Moreover, the reduced data stream, containing only meaningful values and their frequencies, can be processed by automated methods in considerably shorter time and can be easier to understand for human experts.

Conclusions

Automated data analysis and medical diagnosis of large data streams is a complex issue due to performance of computer systems and methods for high-dimensional data (so-called curse of dimensionality).

Instance selection based on thresholds and bounds may reduce data stream by over 90% without losing anomalies which was proven in this paper.

Nonetheless, further studies should be applied. They may be associated with application supervised instance selection techniques and the use of classification methods in terms of selected data, as well as instance selection from different sources [38].

***Authors**: prof. dr hab. inż. Liliana Byczkowska-Lipińska,, University of Computer Sciences and Skills, ul. Rzgowska 17 a, 93-008 Lodz, Poland e-mail: liliana.byczkowska-lipinska@p.lodz.pl*
dr inż. Agnieszka Wosiak, Lodz University of Technology, Institute of Information Technology, ul. Wólczańska 215, 90-924 Lodz, Poland, e-mail: agnieszka.wosiak@p.lodz.pl

REFERENCES

[1] Bellman, R. (2013). Dynamic programming. Courier Corporation.
[2] Bellman, R. E. (2015). Adaptive control processes: a guided tour. Princeton University Press.
[3] Keogh, E., Mueen, A. (2011). Curse of dimensionality. In Encyclopedia of Machine Learning (pp. 257-258). Springer US.
[4] Chen, L. (2009). Curse of dimensionality. In Encyclopedia of Database Systems (pp. 545-546). Springer US.
[5] Abdi, H., Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), pp. 433-459.
[6] Gorban, A. N., Kégl, B., Wunsch, D. C., & Zinovyev, A. Y. (Eds.). (2008). Principal manifolds for data visualization and dimension reduction, Vol. 58, pp. 96-130. Berlin-Heidelberg: Springer.
[7] Byczkowska-Lipińska L., Wosiak A. (2015). Feature Selection and Classification Techniques in the Assessment of the State for Large Power Transformers. Przegląd Elektrotechniczny, R. 91 NR 1/2015, doi:10.15199/48.2015.01.39
[8] Liu L., Özsu M. T. (Eds.) (2009). Encyclopedia of database systems. Berlin, Heidelberg, Germany: Springer. De
[9] Choudhury, M., Lin, Y. R., Sundaram, H., Candan, K. S., Xie, L., & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media?. ICWSM, 10, pp. 34-41.
[10] Holmes, A. (2012). Hadoop in practice. Manning Publications Co..
[11] Buza K., Nanopoulos A., Schmidt-Thieme L., Koller J. (2011, July). Fast classification of electrocardiograph signals via instance selection. In Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on, pp. 9 – 16.
[12] Ramírez-Gallego S., Krawczyk B., García S., Woźniak M., Herrera F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. Neurocomputing, 239, pp. 39 – 57.
[13] García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M., Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), 9.
[14] Pyle, D. (1999). Data preparation for data mining (Vol. 1). Morgan Kaufmann.
[15] Hall, M. A. (1999). Correlation-based feature selection for machine learning.
[16] Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, vol. 3, pp. 1157-1182.
[17] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28.
[18] Giráldez, R. (2005, June). Feature influence for evolutionary learning. In Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation, pp. 1139-1145. ACM.
[19] Kim, J. O., Mueller, C. W. (1978). Factor analysis: Statistical methods and practical issues, Vol. 14. Sage.
[20] Dunteman, G. H. (1989). Principal components analysis. Vol. 69. Sage.
[21] Smith, L. I. (2002). A tutorial on principal components analysis. Cornell University, USA, vol. 51(52), no 65.
[22] Groth, D., Hartmann, S., Klie, S., Selbig, J. (2013). Principal components analysis. Computational Toxicology: Volume II, pp. 527-547.
[23] Bressan, M., Vitria, J. (2003). On the selection and classification of independent features. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(10), pp. 1312-1317.
[24] Liu, H., Motoda, H. (2002). On issues of instance selection. Data Mining and Knowledge Discovery, 6(2), pp. 115-130.
[25] Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., & Kittler, J. (2010). A review of instance selection methods. Artificial Intelligence Review, vol. 34(2), pp. 133-143.
[26] Garcia, S., Derrac, J., Cano, J., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34(3), pp. 417-435.
[27] García, S., Luengo, J., Herrera, F. (2015). Data preprocessing in data mining (pp. 59-139). New York: Springer.
[28] Garcia, S., Derrac, J., Cano, J., & Herrera, F. (2012). Prototype selection for nearest neighbor classification: Taxonomy and empirical study. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34(3), pp. 417-435.
[29] Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine learning, vol. 6(1), pp. 37-66.
[30] Salganicoff, M. (1993, December). Density-adaptive learning and forgetting. In Proceedings of the Tenth International Conference on Machine Learning (Vol. 3, pp. 276-283).
[31] Klinkenberg, R. (2004). Learning drifting concepts: Example selection vs. example weighting. Intelligent Data Analysis, vol. 8(3), pp. 281-300.
[33] Brighton, H., & Mellish, C. (2002). Advances in instance selection for instance-based learning algorithms. Data mining and knowledge discovery, vol. 6(2), pp. 153-172.
[34] Goldberger A.L., Amaral L.A.N., Glass L., Hausdorff J.M., Ivanov P.Ch., Mark R.G., Mietus J.E., Moody G.B., Peng C.K., Stanley H.E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation vol. 101(23), pp. e215-e220, DOI: 10.1161/01.CIR.101.23.e215
[35] Goldsmith R.L., Bigger J.T., Bloomfield D.M., Krum H., Steinman R.C., Sackner-Bernstein J., Packer M. Long-term carvedilol therapy increases parasympathetic nervous system activity in chronic congestive heart failure. American Journal of Cardiology 1997; vol. 80, pp. 1101-1104.
[36] Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., & Kittler, J. (2010). A review of instance selection methods. Artificial Intelligence Review, vol. 34(2), pp. 133-143.
[37] Witten I. H., Frank E., Hall M. A., Pal C. J. (2017). Data Mining: Practical machine learning tools and techniques. Fourth Edition. Morgan Kaufmann.
[38] Guo, L., Chen, F., Gao, C., & Xiong, W. (2012). Performance Measurement Model of Multi-Source Data Fusion Based on Network Situation Awareness. Przegląd Elektrotechniczny, vol. 88(7b), pp. 315-319.