Addressing Challenges in Healthcare Big Data Analytics

Santosh Tirunagari[1], Senthilkumar Mohan[2*], David Windridge[1], and Yashaswini Balla[3]

[1] Department of Computer Science, Middlesex University, The Burroughs, London, United Kingdom
{s.tirunagari, d.windridge}@mdx.ac.uk
[2] School of Information Technology and Engineering, Vellore Institute of Technology, Vellore, India
senthilkumar.mohan@vit.ac.in
[3] Neurosciences Department, Alder Hey Children's NHS Foundation Trust, Liverpool, United Kingdom
yashaswiniballa@doctors.org.uk

Abstract. The exponential growth of healthcare data poses significant challenges for clinical researchers who strive to identify meaningful patterns and correlations. The complexity of this data arises from its high dimensionality, sparsity, inaccuracy, incompleteness, longitudinality, and heterogeneity. While conventional pattern recognition algorithms can partially address issues related to high dimensionality, sparsity, inaccuracy, and longitudinality, the problems of incompleteness and heterogeneity remain a persistent challenge, particularly when analyzing electronic health records (EHRs). EHRs often encompass diverse data types, such as clinical notes (text), blood pressure readings (longitudinal numerical data), MR scans (images), and DCE-MRIs (longitudinal video data), and may only include a subset of data for each patient at any given time interval. To tackle these challenges, we propose a kernel-based framework as the most suitable approach for handling heterogeneous data formats by representing them as matrices of equal terms. Our research endeavours to develop methodologies within this framework to construct a decision support system (DSS). To achieve this, we advocate for the incorporation of preprocessing mechanisms to address the challenges of incompleteness and heterogeneity prior to integration into the kernel framework.

Keywords: Electronic health records · Kernel methods · Bigdata.

1 Introduction

The implementation of electronic health records (EHRs) has accelerated greatly in recent years, resulting in vast amounts of patient data being stored online and easily accessible to healthcare professionals [12, 31, 11]. EHRs enable sharing of patient data across various healthcare settings, necessitating linking of individual patient records from different sources [13]. While the availability of EHRs presents new opportunities for big healthcare analytics, it also poses significant challenges [18, 21].

Electronic health records (EHRs) present a unique set of challenges due to the diverse nature of data they contain, as well as their longitudinal and often incomplete nature [18]. This means that EHRs can include a range of data types, from handwritten notes, EEG signals [22] to medical images and videos [30], which may not be available for every patient or at every time interval. Additionally, EHRs contain longitudinal aspects of patient records, which means that they can include data taken over a period of several years or hours. However, missing values are common in EHRs due to irregular recording intervals and missed appointments [29, 28]. All of these factors can create challenges when analyzing and interpreting EHR data.

Using traditional methods to analyze this data can result in oversimplified conclusions. To help doctors make informed decisions, an effective framework that can sort the aforementioned challenges. The current study advocates developing a kernel-based framework for analyzing EHRs, so doctors can have an efficient clinical DSS to help them make the best possible decisions for their patients.

2 Background

In this section, we provide a brief summary of existing methods in three areas. In Fold 2.1, we discuss techniques for managing data that is collected over time (longitudinal). In Fold 2.2, we review techniques for handling different types of data (heterogeneity). Lastly, in Fold 3.4, we explore methods for dealing with missing data.

2.1 Handling Longitudinal Data

Clinicians use data collected over time to help diagnose and choose treatments for their patients. This data can include information like blood pressure readings, GFRs, HbA1C readings over several years, and videos taken over a period of time. Researchers have developed various methods for analyzing this type of data, some of which can also predict disease progression, even when some data is missing.[33, 23].

2.2 Handling Heterogeneous Data

Considerable research has been devoted to addressing the challenges of handling heterogeneous data using multi-source learning. The literature can be broadly categorized into three types based on research methods [34].

Additive Classifier Models This approach uses machine learning algorithms to learn patterns in different types of data, which are then combined using techniques like bagging or boosting. This method has been used in electronic health records to validate gene expression and protein-protein interaction data [7, 15]. By combining multiple datasets, this approach can make more accurate predictions by finding the overlap between different sets of data.

Graph Models Graph models are used to identify connections between different types of data and represent them as graphs. For example, gene expression data can be represented as graphs to identify protein-protein interactions and sequencing similarities. Graph algorithms can then be used to extract important information from these graphs. Nakaya *et al.*'s study used graph models to identify clusters of genes with shared similarities across different data sources [16]. Bayesian networks and Markov models are examples of specific graph models used in this type of analysis.

Fusion Techniques Two methods of combining data sources are Bayesian fusion and kernel fusion. In Bayesian fusion, each data source is turned into a conditional probabilistic model and then combined through Bayesian networks[20]. For example, Deng*et al* [6] used this method to predict protein-protein interactions and functions using three types of data. However, one downside of Bayesian fusion is that it discards training data during the prediction phase.

On the other hand, in kernel fusion, each data source is represented using matrices called kernels and then combined through linear combinations. This method has the advantage of considering the training data during the prediction phase.

3 Addressing Challenges through Kernel Framework

In this study, we propose utilizing the kernel fusion techniques depicted in Figure 1 due to two primary reasons:

- All data formats can be represented using Mercer Kernels.
- A Mercer Kernel can be expressed as a linear combination of a group of Mercer Kernels.

[scale=0.4]framework.png

Fig. 1. The suggested framework for big data analysis in healthcare.

3.1 Mercer Kernels for different data types

Kernel methods are a type of machine learning technique used to classify and analyze data. They use a mathematical tool called a kernel, which is a matrix that allows for comparisons between pairs of objects in a dataset. By transforming data into a higher-dimensional space through non-linear maps, kernel methods make it easier to classify and analyze data that may be difficult to work with in lower dimensions [4, 25]. This is because the mapping helps to transform data into a format that is easier to work with, making classification and regression operations possible. The most widely used classifier in kernel methods is Support Vector Machines (SVM). One of the primary benefits of kernel methods is that they do not require the coordinates of the higher-dimensional space to be explicitly computed, only the kernel matrix of intra-object comparisons is needed.

Mercer kernels are a type of mathematical function which satisfy Mercer's properties [14, 35] that can be used to analyze different types of healthcare data. They have properties that make them suitable for a variety of formats, such as text, graphs, sequences, shapes, and real numbers. For example, clinical notes and patient information can be handled using NLP parse-tree kernels/LSA kernels [3, 1], while string kernels and random walk kernels are useful for analyzing gene expression data [10, 2]. Edit distance kernels [5] are useful for analyzing shapes, while dot product kernels and polynomial kernels [27] can be used for real number data like blood pressure and glucose levels. With the use of efficient match kernels and pyramid match kernels [9], even sets of pixels/voxels can be processed, such as renal perfusion quantification in DCE-MRI. Fisher kernels [17] are another useful tool for dealing with stochastic data. By using kernel methods, most of the medical data archives can potentially be represented in a kernelized form.

3.2 Linear combination of Mercer Kernels

When combining multiple Mercer kernels, it is possible to create a new Mercer kernel through linear combination while taking care of the kernel weighting problem [8]. This allows for combining different kernels that handle various medical data formats, leading to more linear separability in the data. Additionally, 'meta' kernels can be constructed from these kernels, which further extends the range of possible representations.

To combine multiple kernels, two categories of linear combination methods exist: non-weighted sum and weighted sum methods. The former involves adding or averaging the kernels, while the latter involves linearly parameterizing the combination function using kernel weights. Specifically, given p Mercer kernels $k_1, \ldots, k_p$, the combined kernel function can be defined as:

$$k_\eta(\mathbf{x}_i, \mathbf{x}_j) = f_\eta(\{\mathbf{x}_i^m, \mathbf{x}_j^m\}_{m=1}^p|\eta) = \sum_{m=1}^{p} \eta_m k_m(\mathbf{x_i^m}, \mathbf{x}_j^m) \tag{1}$$

where $\eta = (\eta_1, \ldots, \eta_p)$ are the kernel weights, and $\mathbf{x}^m$ denotes the data in the mth format. We can also define the corresponding feature space mapping $\phi_\eta(\mathbf{x})$ as:

$$\phi_\eta(\mathbf{x}) = \begin{pmatrix} \sqrt{\eta_1}\phi_1(\mathbf{x}^1) \\ \sqrt{\eta_2}\phi_2(\mathbf{x}^2) \\ \vdots \\ \sqrt{\eta_p}\phi_p(\mathbf{x}^p) \end{pmatrix}. \tag{2}$$

where ϕ_m is the feature space mapping corresponding to kernel k_m. The combined kernel function can then be computed in terms of the dot product in the combined feature space:

$$< \phi_\eta(\mathbf{x}_i), \phi_\eta(\mathbf{x}_j) >= \begin{pmatrix} \sqrt{\eta_1}\phi_1(\mathbf{x}_i^1) \\ \sqrt{\eta_2}\phi_2(\mathbf{x}_i^2) \\ \vdots \\ \sqrt{\eta_p}\phi_p(\mathbf{x}_i^p) \end{pmatrix}^T \begin{pmatrix} \sqrt{\eta_1}\phi_1(\mathbf{x}_j^1) \\ \sqrt{\eta_2}\phi_2(\mathbf{x}_j^2) \\ \vdots \\ \sqrt{\eta_p}\phi_p(\mathbf{x}_j^p) \end{pmatrix} = \sum_{m=1}^{p} \eta_m k_m(\mathbf{x}_i^m), \mathbf{x}_j^m) \tag{3}$$

Non-linear combinations using exponentiation or multiplication can also be considered as alternatives to linear combinations.

3.3 Advantages of the Kernel Framework

Using the kernel framework offers several benefits:

Keeps the heterogeneity (diversity) of data intact : The kernel trick transforms different types of data into kernel matrices of equal sizes, which allows for easy integration using linear combination methods. This ensures that the original data's diversity is preserved.

Openness, flexibility, and extendability : The linear combination of kernel matrices allows for the effortless addition or removal of various data types. Additionally, the compatibility of SVMs in statistical modelling makes it possible to use these algorithms in the kernel framework.

Ability to solve large convex optimization problems : Many algorithms have been developed over the past two decades to solve convex optimization problems efficiently. These include interior point methods [24], which make it possible to solve kernel convex optimization problems on a large scale.

3.4 Dealing with Missing Values

When working with a kernel-based framework, there are two main things to consider when dealing with missing data: (1) dealing with missing values over time (longitudinal)in the data and (2) handling missing types of data when integrating different kernels.

Missing values in longitudinal data Electronic health records (EHRs) may have missing information due to irregular recording intervals, such as when blood pressure is measured only during appointments. This can result in missing data due to appointment cancellations or rescheduling. One solution to handle missing numerical data in EHRs is through kernel methods, specifically Gaussian processes (GP). Gaussian processes can predict patient outcomes [26] and interpolate or extrapolate missing data in longitudinal data. By using Gaussian process regression, the best linear unbiased prediction of extrapolation points can be produced, making it useful for modelling time-series data. In a medical context, Gaussian processes can help predict disease progression indicators, such as tumour size, which is a primary outcome variable.

Missing values in heterogeneous data Kernel methods are useful for combining different types of data, but they can encounter a problem when some of the data is missing. This is known as the missing inter-modal data problem, which means that some data in one modality may be missing for certain individuals. This problem is not easily solvable with traditional statistical or machine learning algorithms, but a method called "neutral point substitution" has been developed by Windridge *et al* [32] to address this challenge. This method replaces missing values with a neutral placeholder object in a multi-modal SVM combination. It is more tractable than other state-of-the-art methods that use imputation or naive Bayes fusion [19]. The neutral point substitution method can help overcome the challenges posed by big medical data by providing the necessary expertise and tools.

4 Conclusion

This study has highlighted the challenges faced by big healthcare analytics due to the increasing volumes of electronic health records (EHRs) in the healthcare industry. However, overcoming these challenges could present new opportunities for efficient DSS. To address these challenges, we propose a kernel-based framework that accommodates a variety of heterogeneous data present in EHRs represented as Mercer Kernels at equal terms.

To deal with heterogeneous data, we suggest using existing kernel methods independently for each of the data types and then combining them linearly to form a single kernel matrix effectively. The missing modality problem can be solved using methods such as neutral point substitution (NPS). For longitudinal numerical data with missing values, GP-based regression methods can be employed.

In conclusion, this kernel-based framework could prove to be useful in tackling the challenges faced by big healthcare analytics. It can enable more efficient DSS by effectively accommodating heterogeneous data present in EHRs. Additionally, by using existing kernel methods independently for each data type and combining them linearly, we can obtain a single Mercer kernel matrix that can be used for various machine-learning tasks. Overall, this framework can provide a promising path towards improved healthcare analytics.

References

1. Aseervatham, S.: A local latent semantic analysis-based kernel for document similarities. In: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. pp. 214–219. IEEE (2008)
2. Borgwardt, K.M., Kriegel, H.P.: Shortest-path kernels on graphs. In: Data Mining, Fifth IEEE International Conference on. pp. 8–pp. IEEE (2005)
3. Collins, M., Duffy, N.: Convolution kernels for natural language. In: Advances in neural information processing systems. pp. 625–632 (2001)
4. Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines and other kernel-based learning methods. Cambridge university press (2000)
5. Daliri, M.R., Torre, V.: Shape recognition based on kernel-edit distance. Computer Vision and Image Understanding **114**(10), 1097–1103 (2010)

6. Deng, M., Sun, F., Chen, T.: Assessment of the reliability of protein–protein interactions and protein function prediction. In: Pac. Symp. Biocomputing (PSB 2003). pp. 140–151 (2002)
7. Ge, H., Liu, Z., Church, G.M., Vidal, M.: Correlation between transcriptome and interactome mapping data from saccharomyces cerevisiae. Nature genetics **29**(4), 482–486 (2001)
8. Gönen, M., Alpaydın, E.: Multiple kernel learning algorithms. The Journal of Machine Learning Research **12**, 2211–2268 (2011)
9. Grauman, K., Darrell, T.: The pyramid match kernel: Efficient learning with sets of features. The Journal of Machine Learning Research **8**, 725–760 (2007)
10. Hofmann, T., Schölkopf, B., Smola, A.J.: A review of kernel methods in machine learning. Mac-Planck-Institut für biologische, Kybernetik, Tech. Rep **156** (2006)
11. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C., Verspoor, K.: Biomedical text mining: State-of-the-art, open problems and future challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, pp. 271–300. Springer (2014)
12. Krebs, K., Milani, L.: Harnessing the power of electronic health records and genomics for drug discovery. Annual Review of Pharmacology and Toxicology **63**, 65–76 (2023)
13. de Lusignan, S., Navarro, R., Chan, T., Parry, G., Dent-Brown, K., Kendrick, T.: Detecting referral and selection bias by the anonymous linkage of practice, hospital and clinic data using secure and private record linkage (saprel): case study from the evaluation of the improved access to psychological therapy (iapt) service. BMC medical informatics and decision making **11**(1), 61 (2011)
14. Lyu, S.: Mercer kernels for object recognition with local features. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 2, pp. 223–229. IEEE (2005)
15. Mrowka, R., Liebermeister, W., Holste, D.: Does mapping reveal correlation between gene expression and protein–protein interaction? Nature genetics **33**(1), 15–16 (2003)
16. Nakaya, A., Goto, S., Kanehisa, M.: Extraction of correlated gene clusters by multiple graph comparison. GENOME INFORMATICS SERIES pp. 44–53 (2001)
17. Nicotra, L., Micheli, A., Starita, A.: Fisher kernel for tree structured data. In: Proceedings of the IEEE International Joint Conference on Neural Networks. pp. 1917–1922. Citeseer (2004)
18. Nwegbu, N., Tirunagari, S., Windridge, D.: A novel kernel based approach to arbitrary length symbolic data with application to type 2 diabetes risk. Scientific Reports **12**(1), 4985 (2022)
19. Panov, M., Tatarchuk, A., Mottl, V., Windridge, D.: A modified neutral point method for kernel-based fusion of pattern-recognition modalities with incomplete data sets. In: Multiple Classifier Systems, pp. 126–136. Springer (2011)
20. Poh, N., Merati, A., Kittler, J.: Heterogeneous information fusion: A novel fusion paradigm for biometric systems. In: Biometrics (IJCB), 2011 International Joint Conference on. pp. 1–8. IEEE (2011)
21. Poh, N., Tirunagari, S., Windridge, D.: Challenges in designing an online healthcare platform for personalised patient analytics. In: 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD). pp. 1–6. IEEE (2014)
22. Ramanna, S., Tirunagari, S., Windridge, D.: Epileptic seizure detection using constrained singular spectrum analysis and 1d-local binary patterns. Health and Technology pp. 1–11 (2020)
23. Ribas Ripoll, V.J., et al.: On the intelligent management of sepsis in the intensive care unit (2012)
24. Roos, C., Terlaky, T., Vial, J.P.: Interior point methods for linear optimization. Springer (2006)

25. Scholkopf, B., Smola, A.J.: Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press (2001)
26. Shen, Y., Jin, R., Dou, D., Chowdhury, N.A., Sun, J., Piniewski, B., Kil, D.: Socialized gaussian process model for human behavior prediction in a health social network. In: ICDM. vol. 12, pp. 1110–1115. Citeseer (2012)
27. Smola, A.J., Ovari, Z.L., Williamson, R.C.: Regularization with dot-product kernels. Advances in Neural Information Processing Systems pp. 308–314 (2001)
28. Tirunagari, S., Bull, S., Poh, N.: Automatic classification of irregularly sampled time series with unequal lengths: A case study on estimated glomerular filtration rate. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2016)
29. Tirunagari, S., Bull, S.C., Vehtari, A., Farmer, C., De Lusignan, S., Poh, N.: Automatic detection of acute kidney injury episodes from primary care data. In: 2016 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1–6. IEEE (2016)
30. Tirunagari, S., Poh, N., Wells, K., Bober, M., Gorden, I., Windridge, D.: Movement correction in dce-mri through windowed and reconstruction dynamic mode decomposition. Machine Vision and Applications **28**, 393–407 (2017)
31. Windridge, D., Bober, M.: A kernel-based framework for medical big-data analytics. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, pp. 197–208. Springer (2014)
32. Windridge, D., Mottl, V., Tatarchuk, A., Eliseyev, A.: The neutral point method for kernel-based combination of disjoint training data in multi-modal pattern recognition. In: Multiple Classifier Systems, pp. 13–21. Springer (2007)
33. Yarkiner, Z., Hunter, G., O'Neil, R., de Lusignan, S.: Applications of mixed models for investigating progression of chronic disease in a longitudinal dataset of patient records from general practice. J Biomet Biostat S **9**, 2 (2013)
34. Yu, S., Tranchevent, L.C., Moor, B., Moreau, Y.: Kernel-based data fusion for machine learning: methods and applications in bioinformatics and text mining, vol. 345. Springer (2011)
35. Zhou, D.X.: The covering number in learning theory. Journal of Complexity **18**(3), 739–767 (2002)