CMBES/SCGB

2016 CMBEC39 Conference
Calgary AB
May 24–27, 2016

ANALYSIS OF BIG DATA IN RUNNING BIOMECHANICS: APPLICATION OF MULTIVARIATE ANALYSIS AND MACHINE LEARNING METHODS

Angkoon Phinyomark, Sean T. Osis, Reed Ferber
Faculty of Kinesiology, University of Calgary, Canada

INTRODUCTION

Much of the biomechanical research over the past 20 years has investigated the influence of potential injury risk factors in isolation [1]. More likely, multiple biomechanical and clinical variables interact with one another and operate as combined risk factors to the point that traditional biomechanical analysis methods [2-3] cannot capture the complexity of these relationships. Multivariate analysis and machine learning methods are necessary to identify these complex associations. However, to build accurate classification models, an adequate number of samples are needed, which grows exponentially with the number of features used in the analysis. Therefore, to directly meet this need we have developed the infrastructure and established a worldwide and growing network of clinical and research partners all linked through the world's first automated 3-dimensional (3D) biomechanical gait data collection system: 3D GAIT. Considering that traditional data analytics may not be able to handle these large volumes of data [4], appropriate "big data" analysis methods must be developed [4]. This paper begins with an introduction to our 3D GAIT, followed by an overview of a big data problem in running biomechanics. Next, a comprehensive overview of our proposed and existing methods [2-3,5-8] on the role of big data analytics is presented.

3D DATA COLLECTION SYSTEM

The 3D GAIT system is a deployed turnkey motion capture platform specifically designed for gait analysis using a treadmill. The overall system design is a nexus of 3 main principles: (1) ease-of-use/automation, (2) biomechanics best practices, and (3) data science best practices. Consequently, the system uses off-the-shelf passive motion capture technology, consisting of between 3 and 6 infrared cameras (Vicon Motion Systems, Oxford, UK) along with spherical retroreflective markers that are pre-configured for ease of placement on the subject. Rigid clusters of markers are strapped to the subject's thighs, shanks and pelvis, and markers are taped to the shoes in groupings to define foot movement. During a treadmill session, the cameras operate at 200 Hz for 30 seconds, collecting approximately 150,000 data points representing the 3D coordinates of each marker. These marker data are transformed using rigid-body kinematics [9] into joint angles, which are 3D representations of body movements between segments, over time.

Joint angles from treadmill gait represent a set of non-independent, time-series waveforms, and there are several types of analyses that can be undertaken. In terms of biomechanics best practices, it is considered appropriate to determine a "characteristic" pattern of motion that is representative of the movements for a given subject. Therefore, the 3D GAIT system derives a "characteristic" pattern from a spatio-temporal normalized set of gait cycles, which are segmented using a machine learning approach to account for inter-subject variability in technique [10]. These normalized gait cycles can then be analyzed by: 1) collapsing into a single representative time-series data set by various averaging techniques, and 2) extracting discrete features from each cycle separately, and merging into a representative feature set for a given subject.

After processing, and according to best practices in data science, the final data set is anonymized and packaged for transport. Marker data from the motion capture system, along with biomechanical feature vectors and demographic information (i.e., height, mass,

age, etc.) are securely transmitted via end-to-end encryption to the central server for further processing and storage in a database. These aggregate data, along with critical yet anonymized subject characteristics, allow the potential to statistically model lower limb injury and disease outside of the laboratory setting. More importantly, all data from each centre are automatically entered into a research database. This growing worldwide network currently consisting of 15 research and 50 clinic partners.

BIG DATA PROBLEM

Analysis of running biomechanical data can be considered a big data problem, in the light of the "5V" definition [11], as follows:

1) ***Volume*** (quantity of data): Traditional biomechanical analysis generally involves only a few variables and low subject numbers. Recent biomechanical research, however, the number of variables has increased to ~50-150 discrete variables [2-3], several hundred to thousand variables for joint angle time-series data [5,7], and several thousand to hundred thousand variables for marker coordinate time-series data [12-13]. While most of these studies continue to involve only a small cohort of subjects (10-30) in the analysis [14-15], our database can provide a large cohort of subjects (e.g. 400-500 subjects [2]).

2) ***Variety*** (different data categories): Recent biomechanical analysis involves data from motion capture and also wearable sensors and clinical data: self-reports and lab exams. These data would include continuous, discrete, and categorical data and thus sophisticated statistical methods need to be employed.

3) ***Velocity*** (fast generation of new data): Running related-injuries are often chronic in nature and rehabilitation often takes weeks-to-months. In order to monitor the progress of a rehabilitation program, gait data are generally collected at baseline, and some data are collected once a week over several weeks of the program. On average, 25 new patients are added each week to our database, and 12-15 new clinic partners are added each year.

4) ***Veracity*** (quality of data): Although in general, there is a large divide between clinical research and clinical practice. Since the same data collection system is implemented in both a laboratory and clinical setting, data from motion capture in our database are generally of high quality. However, there is the possibility to have incomplete clinical data (self-reports and lab exams). Fortunately, big data analytics can handle incomplete data sets when necessary.

5) ***Value*** (in the big data): Although the potential value associated with these complex and large data is very high, the real value of big data analytics in running biomechanics still remains to be proven. Much more sophisticated analytics, which incorporate *a priori* knowledge are necessary. In addition, multivariate analysis and machine learning methods could potentially be utilized as an automated system for detecting gait changes related to injury.

DIMENSIONALITY REDUCTION

Initial features

Most investigations of running biomechanics are based kinematic data and have focused on events of the gait waveform such as angles at touchdown and toe-off. Descriptive statistics such as peak angles and excursion are commonly extracted from the gait waveform as well. However, traditional approaches call for the *a priori* selection of features, which relies on sufficient background knowledge and/or subjective opinion. In traditional analysis methods, a large portion of the kinematic data is discarded, which may contain meaningful information related to the between-group differences. While traditional approaches have analysed each joint motion separately, recently the full data set - either a set of representative variables across joints and planes of motion [2-3] or the entire running waveform [5,7] - have been employed as the initial features. However, the dimensionality of the initial features used in the analysis should be carefully chosen due to the fact that several dimensionality reduction methods require an adequate number of samples to obtain stable results. For example, Barrett and Kline [16] recommended that the number of subjects should be at least 50 for a principal component analysis (PCA) method. Unfortunately, initial research involving big data methods have involved small cohort of subjects (10-30) [14-15]. Thus, to minimize the high-dimensionality of the data, big data in terms of

big volume (a large cohort of subjects) is needed, which can be fulfilled by our database.

Feature selection

Instead of choosing the appropriate features based on investigator's background knowledge, feature selection approaches return a subset of the original features using a combination of a search strategy with an objective function. The simplest and most popular approach is to select features with the highest relevance to the target class [2-3,5]. There are two types of measure to score features: filter and wrapper. Wrapper methods use a specific classifier with a cross-validation method to provide a score or the classification rate [6,17] for each subset. Although wrapper methods provide the best performing feature set for a specific classifier, there is no guarantee that this feature set will perform the best for other classifiers. Moreover, the computational cost of wrapper methods is higher than filter methods. To perform wrapper methods for big data, a parallel computing version of cross-validation may be necessary [4]. In contrast, filter methods use interclass distance, or information-theoretic measures, to provide a score. Measures in this field include the effect size [2-3] and the scores of significant tests. Although mutual information has not been applied in this field [18], this measure offers some potential when initial features consist of both categorical data and continuous/discrete data. While filter methods generally provide lower prediction performance than wrapper methods, a selected feature subset is more general and so it is useful for exposing the associations between features. Filter methods can also be used as a preprocessing step [5] for feature extraction, allowing this method to obtain stable results when the dimensionality of initial input is high.

For a search technique, a sequential forward selection (SFS) algorithm is one of the most common search procedures by adding features sequentially. This algorithm has achieved good classification performance to select a subset of discrete variables in our investigations [6,17]. However, the sequential algorithms have a tendency to become trapped in local minima, especially when dimensionality is very high. To deal with a higher-dimensional data, algorithms incorporating randomness into their search procedure are needed to escape local minima, e.g. genetic algorithms (GA). Several popular search techniques have also been developed to work in parallel computing and can be used for big data analytics such as parallel GA [4].

Feature extraction

Instead of selecting the original features, feature extraction approaches transform all the existing features into a new lower-dimensional space. The data transformation can be either linear (as in the most popular method in this field, PCA [2-3,7,12-15,17]) or non-linear (e.g. kernel PCA [8] and self-organizing maps, (SOM) [19]). Specifically, for PCA researchers often use only the first few, or lower-order PCs, which are associated with the most dominant movement patterns. For instance, these PCs are useful for identifying differences in running gait patterns between sex- and age-groups [3]. In contrast, intermediate- and higher-order PCs are often associated with subtle movement patterns. Our research has shown that these PCs are useful for identifying changes in biomechanics after a rehabilitation protocol for injured subjects [3]. PCA can also be extended to model data distributions in high-dimensional space by using a kernel trick called "kernel PCA." The performance of this method for identifying sex and age differences in running gait patterns increases as compared to using the linear PCA [8]. However, the computational cost of these non-linear methods (kernel PCA and SOM) is high in comparison to linear methods, and it may cause a problem in big data analysis. Therefore, supervised feature extraction methods, i.e., a linear discriminant analysis (LDA) and its extended versions should be investigated in future work [20].

CLASSIFICATION AND CLUSTERING

After a final feature vector is created, a supervised or unsupervised learning approach is needed to perform the classification or clustering. For classification, the most popular supervised learning method in this field is a support vector machine (SVM) [2-3,6-7,12-13]. SVM builds a model that predicts whether a new subject fits best in one category or the other (a binary linear classifier). SVM can also efficiently perform a non-linear classification as

well as a multiple classification using multiple binary classifiers. However, the linear kernel exhibits better classification performance as compared to non-linear kernels: polynomial and RBF [6]. For a robust model, an LDA classifier is recommended [17,20]. Unlike SVM and LDA, AdaBoost is another classifier wherein the training process performs the implicit feature selection [21]. However, AdaBoost is sensitive to noisy data and outliers. On the other hand, when the target classes are not available, cluster analysis is needed, e.g. to determine if running patterns for healthy subjects could be classified into homogeneous subgroups [7].

CONCLUSION

In recent years, technological advances now provide researchers with large amounts of data, which can be explored for meaningful patterns. However, traditional data analytics cannot handle these large volumes of data. Therefore, we have developed an automated 3D biomechanical gait data collection system and applied various "big data" statistical methods.

ACKNOWLEDGEMENTS

The CIHR and AIHS Fellowships partially funded this research.

REFERENCES

[1] M. Louw and C. Deary, "The biomechanical variables involved in the aetiology of iliotibial band syndrome in distance runners – A systematic review of the literature," *Phys. Ther. Sport*, vol. 15, pp. 64-75, 2014.

[2] A. Phinyomark, B.A. Hettinga, S.T. Osis, and R. Ferber, "Gender and age-related differences in bilateral lower extremity mechanics during treadmill running," *PLoS One*, vol. 9, e105246, 2014.

[3] A. Phinyomark, B.A. Hettinga, S.T. Osis, and R. Ferber, "Do intermediate- and higher-order principal components contain useful information to detect subtle changes in lower extremity biomechanics during running?," *Hum. Mov. Sci*., vol. 44, pp. 91-101, 2015.

[4] C.W. Tsai, C.F. Lai, H.C. Chao, and A.V. Vasilakos, "Big data analytics: A survey," *J. Big Data*, vol. 2, pp. 21, 2015.

[5] A. Phinyomark, S.T. Osis, B.A. Hettinga, R. Leigh, and R. Ferber, "Gender differences in gait kinematics in runners with iliotibial band syndrome," *Scand. J. Med. Sci. Sports*, vol. 25, pp. 744-753, 2015.

[6] R.K. Fukuchi, B.M. Eskofier, M. Duarte, and R. Ferber, "Support vector machines for detecting age-related changes in running kinematics," *J. Biomech*., vol. 44, pp. 540-542, 2011.

[7] A. Phinyomark, S.T. Osis, B.A. Hettinga, and R. Ferber, "Kinematic gait patterns in healthy runners: A hierarchical cluster analysis," *J. Biomech*., vol. 48, pp. 3897-3904, 2015.

[8] A. Phinyomark, S.T. Osis, B.A. Hettinga, and R. Ferber, "Kernel principal component analysis for identification of between-group differences and changes in running gait patterns," *accepted to Proc. XIV Mediterr. Conf. Med. Biol. Eng. Comput*., 2016.

[9] I. Söderkvist and P.A. Wedin, "Determining the movements of the skeleton using well-configured markers," *J. Biomech*., vol. 26, pp. 1473-1477, 1993.

[10] S.T. Osis, B.A. Hettinga, J. Leitch, and R. Ferber, "Predicting timing of foot strike during running, independent of striking technique, using principal component analysis of joint angles," *J. Biomech*., vol. 47, pp. 2786-2789, 2014.

[11] Y. Demchenko, Z. Zhao, P. Grosso, A. Wibisono, and C. de Laat, "Addressing big data challenges for scientific data infrastructure," *Proc. IEEE 4th Int. Conf. Cloud Comput. Technol. Sci.*, pp. 614-617, 2012.

[12] B.M. Eskofier, P. Federolf, P.F. Kugler, and B.M. Nigg, "Marker-based classification of young-elderly gait pattern differences via direct PCA feature extraction and SVMs," *Comput. Methods Biomech. Biomed. Eng*., vol. 16, pp. 435-442, 2013.

[13] C. Maurer, P. Federolf, V. von Tscharner, L. Stirling, and B.M. Nigg, "Discrimination of gender-, speed-, and shoe-dependent movement patterns in runners using full-body kinematics," *Gait Posture*, vol. 36, pp. 40-45, 2012.

[14] C. Maurer, V. von Tscharner, M. Samsom, J. Baltich, and B.M. Nigg, "Extraction of basic movement from whole-body movement, based on gait variability," *Physiol. Rep*., vol. 1, e00049, 2013.

[15] P. Federolf, K. Tecante, and B. Nigg, "A holistic approach to study the temporal variability in gait," *Gait Posture*, vol. 45, pp. 1127-1132, 2012.

[16] P.T. Barrett and P. Kline, "The observation to variable ratio in factor analysis," *Personality Study Group Behav.*, vol. 1, pp. 23-33, 1981.

[17] D. Kobsar, S.T. Osis, B.A. Hettinga, and R. Ferber, "Gait biomechanics and patient-reported function as predictors of response to a hip strengthening exercise intervention in patients with knee osteoarthritis," *PLoS One*, 10, e0139923, 2015.

[18] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinf. Comput. Biol.*, vol. 3, pp. 185-205, 2005.

[19] S. Hoerzer, V. von Tscharner, C. Jacob, and B.M. Nigg, "Defining functional groups based on running kinematics using self-organizing maps and support vector machines," *J. Biomech*., vol. 48, pp. 2072-2079, 2015.

[20] A. Phinyomark, H. Hu, P. Phukpattaranont, and C. Limsakul, "Application of linear discriminant analysis in dimensionality reduction for hand motion classification," *Meas. Sci. Rev*., vol. 12, pp. 82-89, 2012.

[21] B.M. Eskofier, M. Kraus, J.T. Worobets, D.J. Stefanyshyn, and B.M. Nigg, "Pattern classification of kinematic and kinetic running data to distinguish gender, shod/barefoot and injury groups with feature ranking," *Comput. Methods Biomech. Biomed. Eng*., vol. 15, pp. 467-474, 2012.