An Imaging Framework for the Analysis of Longitudinal High-Dimensional Data

Jeffery L. Painter
GlaxoSmithKline, Medical Analytics, RTP, NC, USA

Abstract— *High dimensional data on its own poses several difficulties when applying traditional statistical modeling methods. Adding a time series component imposes an even heftier burden on finding methods capable of dealing with causation and prediction modeling with multivariate data. Often we are faced with the proposition of being able to answer only one question at a time, especially with regards to electronic medical records, and each of these analyses can take days to complete using a method such as Bayesian logistic regression. Here we propose a method for reducing the parameter space of high-dimensional data through imaging in order to enable simultaneous evaluation of multiple parameters. Finally, we evaluate some of the limitations found when attempting to apply some widely used image comparison algorithms to the resulting imaged data.*

Keywords: high-dimensional longitudinal data, image comparison algorithms, electronic medical records, observational data, parameter space reduction, UMLS Metathesaurus

1. Introduction

There is a great deal of interest in being able effectively to analyze large observational data sets. However, the challenges in this arena often stem from the fact that we are typically looking at several different parameters of interest and their interactions over time. There is an information overload problem, not simply a data overload problem. While it is true that the amount of data collected is growing at increasing rates, the complexity of that data is also increasing. Therefore, the value that is to be gained from improvements in the mechanisms for navigating data of high levels of complexity will only increase as time goes on.

It may be fruitful to begin thinking about large data in a different way. This investigation proposes that in order to reduce the high-dimensionality of longitudinal data, the parameter space be shrunk to a 2-D visual representation in order to take advantage of image comparison methods rather than using traditional statistical analysis of multivariate data. While it is not clear whether any benefit may arise from thinking of high-dimensional data in this way, we create a framework for evaluating the merits of this approach and enable the possibility of future research into the effectiveness of imaging longitudinal high-dimensional data as a means for navigating large data sets. In particular, we apply this method to patient data captured in large electronic medical records systems.

2. Framework

This system proposes a framework show in Figure 1 whereby observational data is transcribed into an image representing each patient within a large observational database. Multiple parameters of interest are represented within a single image.

Those images are then compared to one another using an image comparison algorithm to generate a distance metric between two individuals within a single database. Once a distance metric is in place, it can then be used in the generation of a graph represented by a symmetric distance matrix whose nodes consist of individual patients where every edge of the graph is weighted with the distance found between them. The weighted graph can then be employed to explore large observational data sets quickly by using graph theoretic methods.

Consequently computation is relegated simply to walking the graph rather than having to evaluate each patient one condition or drug at a time. The sequence of events relative to each patient is accounted for by mapping the events occurring within each patient across a time axis.

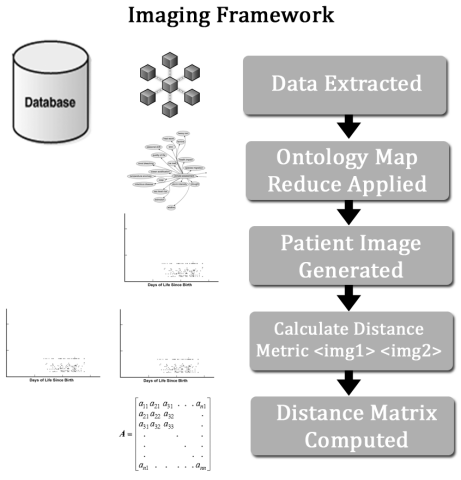

Figure 1: Imaging Framework Overview

2.1 Types of Data

The idea of "information overload" has come to meet the modern investigator head on. As Toffler remarked: "When the individual is plunged into a fast and irregularly changing situation, or a novelty-loaded context ... his predictive accuracy plummets. He can no longer make the reasonably correct assessments on which rational behavior is dependent." (1)

Such is the case with the modern landscape of business intelligence and analytics today. The ultimate vision is that the more data we have at our disposal, the more informed our decision making process will become and the more our predictive accuracy and understanding of why things are the way they are will also improve. These are the goals of modern data mining and statistical methodology; yet delivering those results in real time is still rather difficult and requires huge investments in machinery capable of analyzing vast data by looking for patterns among all the variables of interest.

Woods, Peterson and Roth identified three problems facing the avalanche of data overload: the clutter problem, workload bottlenecks - where there is simply too much to do and too little time to do it in, and finding significance within the data. (2). These are issues still facing us today, and we hope to address each through a novel application of visualizing data. The visualization of patients is not necessarily for the implicit ability to view data by an observer, but to place the issue of comparing images directly to the computer in hopes that with enhanced image analysis techniques we can begin to extract meaningful relations between patients exhibiting observations in complex parameter spaces.

While the impetus driving the research presented here is to find new methods for coping with the growing complexity of large scale observational data as it relates to electronic medical records, insurance claims data and the like, the framework should be extensible to other types of data exhibiting similar characteristics. This includes any data which is collected over time, such as customer preferences, product rating data, purchasing history or more complex applications which track several different, yet related areas of interest over time. Also included might include applications as diverse as manufacturing resource allocation as it contributes to the overall sales performance and product distribution needs of a company over time.

2.2 Multiple Parameters

In the field of electronic medical records (EMR), electronic health records (EHR), insurance claim and private practice data, these systems aim to capture multiple parameters of interest regarding patient health over the time span of coverage for any given patient. Electronic medical record systems are generally more inclusive since it serves as a singular repository for collecting all the information about a single patient over a long period of time. The period of time covered could include the entire life-span of an individual.

Effectively representing this data for patient-to-patient comparison poses many problems. First is the issue of representing multiple parameters effectively. Fodor notes that "traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation" (3).

In large observational data, the drugs that any given patient takes are typically coded in one particular coding scheme while the observations (the medical conditions any particular patient experienced or was observed as having) are recorded in yet a different terminology.

In RxNorm and similar drug ontologies, there are over 14,000 products indicated at the lowest level and medical conditions and observations may be represented by one of over 100,000 values found in the Read[1] terminology for example. Since the average lifespan of a given person could exceed more than 70 years, representing the multitude of potential events, drugs, and medical conditions a person experiences over an entire lifetime seems at first like an impossible task to manage, and yet, we must be capable of comparing this patient to millions of other unique patients within the database who have their own unique combinations of events over time.

There may also be interest in other parameters beyond just drugs and conditions, such as various laboratory values and measurements recorded such as blood pressure, height and weight and other factors that are taken into account over the entire data collection period of any given patient within a health care setting. Similarly, with more complex web applications where several systems are merged together to provide seamless integration of services, the overlap of system services might pose similar data collection issues where several events occur simultaneously, yet possibly unrelated directly to one another across thousands of potential users.

By thinking of each parameter as an individual axis within an image, we can begin to imagine a scheme by which each category of potential values could overlap and give a single composite image consisting of competing parameters over time to represent a patient, a customer or a business process in the real world. The x-axis in each of these images would represent the time component so that as the patient or customer becomes older in the system either new entries are appended onto the image with either the axis fixed at some point in time relative to the individual (such as the date of birth), or all the images generated have a fixed axis.

The advantage to fixing the x-axis for every single person to the date of birth, at least for patient data, is that trends occurring at specific points in time of life will emerge among

[1]The Clinical Terms Version 3 (Read Codes)© are maintained by the (UK) National Health Service Information Authority.

all the patients at similar points in time, with a slight shift to either the left or right, depending on whether a series of events occurs earlier or later in life. If one were to fix the axis in time from a specific event, then this might make more sense for business oriented data where you want to investigate product trends at various points in time, and to use that to help predict future needs more consistently.

2.3 Image Comparison Methods

Once you have the capability to create an image and have addressed the x-axis issue of how to fix the events in time, you can then begin thinking of ways to compare those images. For the initial work presented here, the algorithm *PerceptualDiff* (4) was used to generate a distance measure between any two individual patients.

Image comparison methods at this point are still limited for the application of comparing sparse images as generated from our initial sample of patients.

2.4 Distance Metrics

By using the image comparison tools, we generate a distance metric between any two images. This in turn can be used to generate a graph where each node on the graph represents an observed data point comprised of several multivariate observations over time (a patient in our case) and the distance between those nodes is equal to the output of our image comparison algorithm.

In the case of *PerceptualDiff*, the output generates a pixel count difference which we can use to populate the graph of distance measures between each and every single patient. "These metrics perform signal processing on the two images to be compared, mimicking the response of the human visual system to spatial frequency patterns and calculating a perceptual distance between the two images." (5)

The distance metric is vital to making effective use of the graph generated from image comparison in this type of framework. Any significant statistical analysis imposed over the edges of the graph will depend on the meaningfulness of this metric (6) (7) (8). Therefore, if there are improvements to be gained from future research, it is in finding more advanced image comparators that can generate a distance metric which takes into account various attributes of the imaged data and not simply a pixel by pixel comparison as is done in *PerceptualDiff*.

3. Imaging Longitudinal Data

Our initial effort was to attempt to generate an image representative of a patient or customer taking into account that events happen in a specific order over time. For electronic health records, we are primarily interested in two parameters: the drugs a patient takes and the medical conditions which are observed prior to and after the drug events.

Each of these parameters occupies a space of potential variables which denote a drug or a medical condition. The JFreeChart (9) library was employed to generate an *XYPlot* of each patient utilzing multiple series to represent each parameter space.

JFreeChart was not the initial choice for imaging the data. Initial work was done using the *matplotlib* Python library (10). While *matplotlib* offered many features that were attractive to this type of application, ultimately, the images generated by this library were too large in size for scaling up the framework for production levels. The JFreeChart library allowed for finer grained control over the image creation process, including allowing us to specify transparency levels within the image. By making use of this feature, the images generated from JFreeChart averaged around 9kb, while the images generated from *matplotlib* exceeded 70kb per image.

The size of the image has an enormous impact on the ability of *PerceptualDiff* to calculate its distance metric. The initial images from *matplotlib* took over a minute and a half to calculate the distance between two patients (and required more than 500MB of system memory), while the smaller images generated from JFreeChart could be compared in less than 20 seconds.

3.1 Parallel Spaces

Since each parameter is realized by most patients in our database, it is necessary to include both parameters (drugs and conditions) within a single image. For future work, we would also hope to include even more parameters, such as laboratory values, other health factors (e.g. weight), and so on. Each parameter adds to the number of spaces which must be accommodated within the image itself.

Again, multiple parameters need to be represented on a single axis to aid in reducing the overall image size of generated. As noted in the previous section, the larger the size of the images, the more compute time is required to make a comparison between images.

Inselberg observes that "visualization provides insight through images, and can be considered as a collection of application-specific mappings." (11). He speaks of the Cartesian plane as a mechanism for enabling not just 2 or 3 dimension representations, but allows us to simultaneously map N-dimensions with parallel coordinate systems. Extending this idea, we can create multidimensional mappings in 2-D space by extending not the number of coordinate systems, but rather the representations within a single coordinate system through a color "dimension". In the case of overlapping elements within a single 2-D space, we can represent those points through yet a third or fourth color indicating that there is more than one x event occurring at a particular y-coordinate. For n-dimensions, this can be accomplished through $(2^n - 1)$ number of colors (Mersenne prime number). For example, with two parameters, an overlap would require a third color. For three parameters, you would need three colors to represent the base parameters, three for the occurrence of two simultaneous events and a final color if

all three parameters occur at the same time for a total of seven colors.

For example, let $P = \{a, b, c \ldots\}$ be parameters of interest to graph on parallel axis. Then the number of colors required to represent the possible combinations of simultaneous events on the y-axis follows:

$$C_2 = \{(a, b)(ab)\}$$
$$C_3 = \{(a, b, c), (ab, bc, ac), (abc)\}$$
$$C_4 = \{(a, b, c, d), (ab, ac, ad, bc, bd, cd), (abc, abd, bcd, acd)$$

and subsequently, $|C_2| = 3$, $|C_3| = 7$ and $|C_4| = 15$.

From this pattern, it is observed that there are $M_n = 2^n - 1$ colors required. In order to create a 2-D visualization, we simply add a new series in the output of JFreeChart's *XYPlot* or, in effect, add an additional color to represent multiple parameters within a single image. In the case of multiple parameters occurring within the same point in a single image, those events may be expressed by using yet another color to represent an overlap of parameter spaces.

3.2 Parameter Space Reduction through Ontological Mapping

As noted earlier, the parameters of interest in the patient data we are analyzing is primarily composed of drug exposure data and observed or reported medical conditions over time. When a patient receives a drug prescription, it is typically indicated that the drug be taken over some length of time, and therefore, those become known as "drug eras" (12) which indicate the start and end date of an exposure period for some time duration found in the patient's recorded medical history. The number of possible drugs in this parameter space exceeds 14,000 branded names and strengths.

The point of most traditional statistical analysis has been to distinguish the signal from the noise, addressing the third point of the data overload problem mentioned by Woods (2). However, these methods mostly attempt to look at each parameter one at a time (3). The benefit of applying ontological mappings to the data extracted for each database member is that in essence there is no need to a priori identify those parameters which have little or no impact on predictive value, but rather, those observations which do have meaning should become apparent through the image comparison algorithm. Patterns which reoccur at high frequency should begin to emerge and announce themselves by a relatively lower distance between patients exhibiting similar behavior over time.

However, by applying an ontology such as RxNorm or SNOMED CT's drug and medicament hierarchy (13), many of these individual drugs will collapse from low level drug references into higher level categories which relate similar drugs together in a meaningful way. Similarly, we can apply the same strategy to mapping one of the over 100,000 medical conditions into higher level categories which relate similar concepts to one another.

3.2.1 Proximity Based Positioning of Related Concepts

In the Read coding system, there is a hierarchical structure contained within the coding scheme, but neither the breadth nor depth of this hierarchy is consistent throughout the codes which annotate it. The first image shown is a single patient with multiple drug and condition events recorded starting around the age of 40. The raw data is unorganized, and covers a multitude of medical conditions seen in Figure 2. In the next image, Figure 3, we see the same patient utilizing the MedDRA[2] hierarchy to organize those same conditions into regionalized areas that are closely related. The MedDRA hierarchy serves as our reference system and we mapped the Read codes into it by way of the Unified Medical Language System Metathesaurus[3].

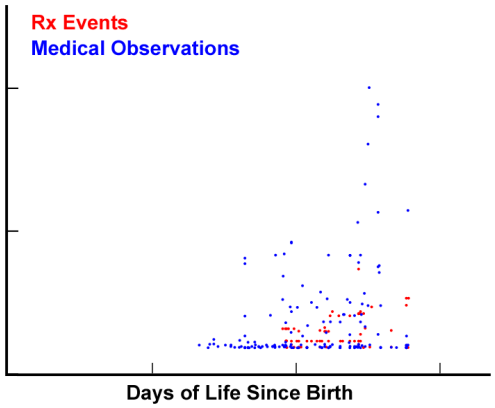

Figure 2: Raw Patient Data

Once the mapping was completed, we were able to make use of the strict 4-level hierarchy (14) found within the MedDRA terminology to collapse each medical condition into a single high level category and also into a secondary group level mapping. This is shown in the final image of the patient in Figure 4. Now it is apparent that the pixels representing medical conditions (in blue) are arranged in a more organized manner as the ontology reduces the overall parameter space from 100,000 or more condition categories to less than 20,000 high level categories and even fewer group level terms.

[2]MedDRA® (Medical Dictionary for Regulatory Activities) is a registered trademark of the International Federation of Pharmaceutical Manufacturers Association

[3]UMLS Metathesaurus is a project of the (US) National Library of Medicine, Department of Health and Human Services. Available at: http://www.nlm.nih.org/research/umls/

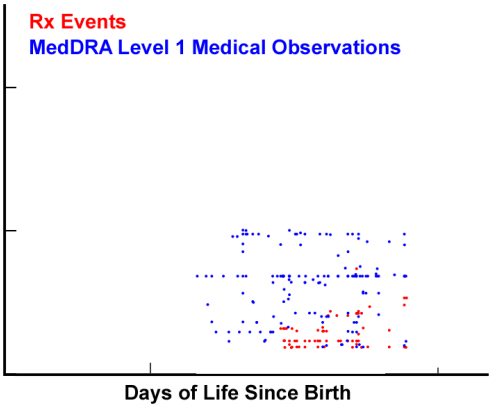

Figure 3: MedDRA Hierarchy Applied to Conditions

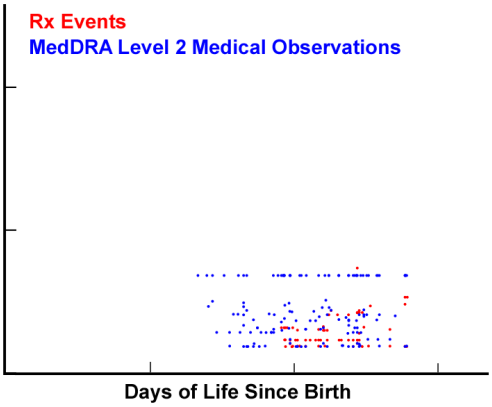

Figure 4: MedDRA Group Level Abstraction

3.3 Captured Time Component

By default, the allocation of events on our X,Y coordinate plane takes into account the longitudinal nature of the data we are evaluating. Many of the statistical methods, such as logistic regression or a priori basket analysis, typically used in epidemiological studies fail to capture the time component. In such methods, we are simply looking at whether or not two events occur within a patient's observed record.

By placing the events sequentially along the X-axis, we expose within the image not only the proximity of closely related concepts, but also put them in their proper context with respect to time. For example, if a patient were to develop high blood pressure followed by a diagnosis of diabetes, we might see at some later point in life a cardiac event. The time component is accounted for by placing each of these events in a sequential ordering. The image comparison will take into account all of these factors when creating a distance measure between any two patients.

4. Image Comparison

Once we have generated the images representing our patients, the next step is to apply an image comparison algorithm in order to calculate a distance measure between each pair of images.

4.1 Calculating Distance Between Images

A preliminary search for image comparator methods revealed that *PerceptualDiff* is widely regarded as one of the best programs to use. It has a simple format, much like the traditional UNIX diff command, and *PerceptualDiff* is capable of taking two images as input and producing as its output a numeric value indicating the number of pixels by which image one differs from image two. A script was written to process several thousand images across a 48 node cluster. The initial sample included around 1,500 patients and took roughly four days to compute the complete graph of distance measures between each and every patient, with a single comparison taking slightly less than 20 seconds. There are obvious computational hurdles to overcome. The *PerceptualDiff* algorithm does not yet support parallel processing to take advantage of multiple cores.

While initially this seemed quite promising, several limitations immediately presented themselves. It is often the case that patients found within large observational data sets will have very sparse records. That is, those patients may have been observed only once or twice, and as a result, the number of events recorded these types of patients may be limited to ten or fewer medical conditions. When taking time into account, it becomes apparent that there is a great deal of "white space" which then appears in a patient's image file.

The distance between these types of patients will be calculated to be very low (depending on the number of conditions), and therefore even a close proximity measure between the two will be effectively meaningless. And so *PerceptualDiff* is misleading for a large number of data sets.

But even traditional analysis methods will have difficulty with these kinds of patients, and while the distance metric may not be very meaningful, it does segment these members of the database quite effectively from the others. Therefore, it does provide some use in that it does help us to navigate large observational data sets.

Similarly, the converse is true. When a single patient exhibits thousands of events and hundreds of drug prescriptions, these individuals are also segmented from the rest of the population. In effect, the image comparison algorithm does help identify individuals who would be considered outliers of the database.

5. Conclusion

The imaging framework described here offers a methodology through which longitudinal high-dimensional data can

be transcribed into an image-based representation to support the application of image comparison algorithms yielding a calculation of distance measures between individual observations found in large observational data sets.

Our methods show that the image generation process can be achieved in linear time; however, the amount of time required in calculating the edge weights for the complete graph yielded from the image comparison algorithm grows exponentially in relation to the total number of data set members. Therefore, the impact of the image comparison algorithm in terms of time complexity is the biggest hurdle faced in employing this methodology.

The ultimate question is how meaningful the calculated distance measure is and whether it can meaningfully be applied to data of this type. Without taking into account time shift, geographical proximity and more advanced probabilistic image comparison metrics, the applicability of the computed distance graph may be severely limited until further improvements in this area are made.

Additionally, it is often the case that individual records (patient, customer or other entity) may only contain a sparse set of observations over a long period of time. Patient data in electronic health records, insurance claims data and related large scale patient records exhibit this property. We may see a single patient out of several million possible patients possessing a patient history that is rather extensive in time, but only exhibits a few characteristic traits over the entire lifespan of observed data points.

Due to the nature of patient data, when an attempt is made to create a representative "image" based on the patient record, we tend to see very sparse images. That is, typically, the image generated contains a large proportion of white space with relatively few pixels indicating an event that actually happened at some point in time. This sparseness creates a problem when we attempt to compare the images using a traditional image comparison algorithm such as *PerceptualDiff*.

Another limitation found with *PerceptualDiff* is that it can only compare one picture to another in a very rigid fashion. It looks for single pixel differentiation between one picture and another without taking into account the proximity of those pixels to one another, or a shift of the image. Analogously, many text differencing programs, such as UNIX diff, are ignorant of contextual information in much the same manner. There may be reoccurring patterns which fail to be discovered due to the inability of *PerceptualDiff* to take into account time shift or relative proximity among recorded observations.

Our future research will focus on improved image comparison algorithms and related research concerning how to apply the graph of distance measures to data mining large observational data sets. In addition, we hope to demonstrate performance metrics of traditional statistical analysis methods in comparison with the image comparison methodology described herein.

6. Acknowledgements

The author would like to thank colleagues Alan Menius (GlaxoSmithKline) and Xiaoshan Li (NCSU Department of Statistics) for contributing to the initial discussions and evaluation of this framework.

References

[1] A. Toffler, *Future Shock*, Random House, 1970.

[2] David D. Woods, Emily S. Patterson, Emilie M. Roth, and Klaus Christoffersen, "Can we ever escape from data overload? a cognitive systems diagnosis", *Cognition, Technology and Work*, vol. 4, pp. 22–36, 2002.

[3] K. Fodor, "A survey of dimension reduction techniques", Tech. Rep. UCRL-ID-148494, Lawrence Livermore National Laboratory, 2002.

[4] "Perceptual image diff", 2012, Available online: http://pdiff.sourceforge.net/.

[5] Hector Yee, Sumanita Pattanaik, and Donald P. Greenberg, "Spatiotemporal sensitivity and visual attention for efficient rendering of dynamic environments", *ACM Trans. Graph.*, vol. 20, no. 1, pp. 39–65, Jan. 2001.

[6] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification", *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[7] Steven C. H. Hoi, Wei Liu, Michael R. Lyu, and Wei-Ying Ma, "Learning distance metrics with contextual constraints for image retrieval", in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, Washington, DC, USA, 2006, CVPR '06, pp. 2072–2078, IEEE Computer Society.

[8] Shang, Ming-Sheng, Lü, Linyuan, Zeng, Wei, Zhang, Yi-Cheng, and Zhou, Tao, "Relevance is more significant than correlation: Information filtering on sparse data", *EPL*, vol. 88, no. 6, pp. 68008, 2009.

[9] "Jfreechart", 2012, Available online: http://www.jfree.org/jfreechart/.

[10] "matplotlib", 2012, Available online: http://matplotlib.sourceforge.net/.

[11] A. Sheth and V. Kashyap, "Parallel coordinates for visualizing multidimensional geometry", in *New Techniques and Technologies for Statistics II*. 1997, pp. 269–271, IOS Press.

[12] Stephanie J Reisinger, Patrick B Ryan, Donald J O'Hara, Gregory E Powell, Jeffery L Painter, Edward N Pattishall, and Jonathan A Morris, "Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases", *Journal of the American Medical Informatics Association*, vol. 17, no. 6, pp. 671–674, November 2010.

[13] G.H. Merrill, P.B. Ryan, and J.L. Painter, "Construction and annotation of a UMLS/SNOMED-based drug ontology for observational pharmacovigilance.", in *Proceedings of the Intelligent Data Analysis for bioMedicine and Pharmacology*, Washington, DC, 2008.

[14] G.H. Merrill, "The MedDRA paradox", in *AMIA Annual Symposium Proc*, Washington, DC, 2008, pp. 470–474.