2012

Multivariate Visual Clustering of Single Nucleotide Polymorphisms and Clinical Predictors using Cherno Faces

Shalem Lee
University of Adelaide

Sharon Lee
University of Queensland

Gus Decker
University of Adelaide

Claire Roberts
University of Adelaide

Publication Details

Shalem Lee, Sharon Lee, Gus Dekker and Claire Roberts, Multivariate Visual Clustering of Single Nucleotide Polymorphisms and Clinical Predictors using Cherno Faces, Proceedings of the Fifth Annual ASEARC Conference - Looking to the future, 2 - 3 February 2012, University of Wollongong.

Multivariate Visual Clustering of Single Nucleotide Polymorphisms and Clinical Predictors using Cherno Faces

Abstract
With advanced technology, collection of health-related data is undertaken on a large scale, producing large and high-dimensional data. Visualization of such data is important and useful for further statistical analyses such as classification and clustering. However, visualizing large multivariate datasets is challenging, especially for high dimensional data, as they are often complex and confounded. Currently, visualization for Single Nucleotide Polymorphisms (SNPs) and clinical predictors of disease are assessed separately. As there is increasing evidence of genetic-environmental interactions for pregnancy complications, prediction models based solely on either clinical measurements or genetic risk factors may be inadequate. Hence, we present an example of multivariate visualization on combinations of clinical measurements and SNPs through Cherno faces, and perform visual clustering for prediction of Preterm births (PTB). A random sample containing 100 patients (Uncomplicated pregnancy= 92, PTB=8) with 11 clinical and 4 genetic predictors are visualized into faces with various style of eyes, ears, nose and hair, showing two groups with similar face characteristics amongst Uncomplicated pregnancies and Preterm births.

Keywords
Multivariate data, Visualization, Cherno faces, Clustering, SNP

Publication Details
Shalem Lee, Sharon Lee, Gus Dekker and Claire Roberts, Multivariate Visual Clustering of Single Nucleotide Polymorphisms and Clinical Predictors using Cherno Faces, Proceedings of the Fifth Annual ASEARC Conference - Looking to the future, 2 - 3 February 2012, University of Wollongong.

Multivariate Visual Clustering of Single Nucleotide Polymorphisms and Clinical Predictors using Chernoff Faces

Shalem Lee*, Sharon Lee†, Gus Dekker*, Claire Roberts*

*Robinson Institute, The University of Adelaide, Adelaide 5005
†School of Mathematics and Physics, The University of Queensland

Abstract

With advanced technology, collection of health-related data is undertaken on a large scale, producing large and high-dimensional data. Visualization of such data is important and useful for further statistical analyses such as classification and clustering. However, visualizing large multivariate datasets is challenging, especially for high dimensional data, as they are often complex and confounded. Currently, visualization for Single Nucleotide Polymorphisms (SNPs) and clinical predictors of disease are assessed separately. As there is increasing evidence of genetic-environmental interactions for pregnancy complications, prediction models based solely on either clinical measurements or genetic risk factors may be inadequate. Hence, we present an example of multivariate visualization on combinations of clinical measurements and SNPs through Chernoff faces, and perform visual clustering for prediction of Preterm births (PTB). A random sample containing 100 patients (Uncomplicated pregnancy= 92, PTB=8) with 11 clinical and 4 genetic predictors are visualized into faces with various style of eyes, ears, nose and hair, showing two groups with similar face characteristics amongst Uncomplicated pregnancies and Preterm births. The faces identified as PTB appear to have either a tall hair style or no ears, which correspond to whether the mother was herself born preterm, and SNPs in TGFβ and IL1β genes.

Key words: Multivariate data, Visualization, Chernoff faces, Clustering, SNP

1. Introduction

Statistical and computational analyses nowadays are much complicated, resulting from the huge amount of high-dimensional data generated. Advanced technologies, such as High-throughput genotyping [1], have not only increased the efficiency of obtaining DNA-sequences, but also vast quantity of data collected. Visualization of such large-scale data has become a great challenge, especially for high-dimensional data, which are often complex and confounded.

In reproductive epidemiology, it is of great interest to investigate possible factors that can contribute to pregnancy complications such as Preterm birth (PTB). Data collected in these studies typically includes a large number of (both categorical and continuous) clinical measurements as well as genetic predictors such as Single Nucleotide Polymorphisms (SNPs), and thus poses a challenge to display the predictors concurrently. As a result, there is a need for simple yet effective methods for visualization of such datasets. Current approaches typically assess SNPs and clinical predictors separately. However, as there is growing evidence of familial tendency and genetic-environmental interactions for pregnancy complications, looking solely at either clinical measurements or genetic risk factors may not be adequate to obtain a complete view of possible contributing predictors for a disease. To address this issue, we propose the use of Chernoff faces for simultaneous visualization of a range of SNPs and clinical measurements, and demonstrate the usefulness of this technique in clustering a subset of the SCOPE pregnancy database.

2. Visualization of SNPs and Clinical Predictors

Single Nucleotide Polymorphisms (SNPs) are the most abundant type of genetic variations of an individual's DNA sequence. Since SNPs can differentiate different inherited forms of a gene, they have been widely used in identifying genetic markers for phenotypes and diseases [2, 3, 4]. This paper mainly focuses on clustering disease based on the heterozygosity or homozygosity of SNPs.

Visualization of SNPs can serve as a quick method to observe the genetic variations in the data obtained. A number of SNP visualization tools have been devel-

*E-mail: shalem.lee@adelaide.edu.au (Shalem Lee)

oped, including SNP-VISTA [5], SNPTools [6], AssociationViewer [7], VizStruct [8]. While many of these tools are useful in providing an in-depth view of the genetic information and gene associations within and between various treatments or disease, it is difficult to obtain an overall view that incorporates clinical predictors, e.g. phenotypes, dietary information or social characteristics.

On the other hand, clinical predictors can be visualized in many ways. These range from histograms to Mosaic plots for categorical predictors, and from scatterplots to 3D density graphics for continuous predictors. Yet, there are only a few multivariate visualization methods available that can incorporate both continuous and categorical variables, especially for high-dimensional data. In this paper, we introduce an interesting 2D visualization tool for assessing SNPs and clinical predictors concurrently using Chernoff faces.

3. Visualization using Chernoff Faces

Introduced by Herman Chernoff in 1971 [9], Chernoff faces is a powerful graphical visualization tool for multi-dimensional data. In simple terms, each variable or dimension of the data is geometrically mapped to a particular feature on a cartoon face according to some mathematical rules. The shape, size and location of ears, eyes, nose and mouth, for example, are controlled by different variables. Among various glyph type displays, Chernoff faces are particularly effective due to our exquisite sensitivity to facial expressions which facilitates easy perception of multiple measurements in parallel.

For p-dimensional data, each variable $X_1, X_2, ..., X_p$ is assigned to a feature on the schematic face and rules are constructed to determine the coordinate, size and curvature of each feature. For example, the first dimension X_1 may corresponds to the height of the face, X_2 determines the face width, X_3 specifies the curvature or structure of the face, and so on. The idea behind this method is that if two data points are very similar, their corresponding faces should appear similar; and if they are very different, this should also reflect through subtle differences in their facial features. Hence, this allows for fast and easy visual comparison of multivariate data.

4. Clustering of Preterm Births

The data used in this study are obtained from the Screening fOr Pregnancy Endpoints (SCOPE) project [10], which aims to build a pregnancy database and biobank to screen candidate markers of pregnancy disease. This database contains comprehensive records of maternal and paternal health conditions, social characteristics, dietary practices, pregnancy history, and details of antenatal visits, together with 100 maternal, paternal and neonatal SNPs.

A random sample of 100 patients were obtained from SCOPE, with 92 uncomplicated pregnancy cases and 8 preterm birth (PTB) cases. Preterm birth is defined as onset of labour before 37 weeks of gestation, and it is one of the major pregnancy complications which occurs in approximately 5-10% of births [11, 12]. Preterm infants are likely to suffer serious morbidities and develop long-term health and developmental problems [13, 14, 15, 16, 17]. Considering the significant long-term effects, an effective method to identify or predict high-risk individuals is essential. A variety of clinical and genetic risk factors have been found to be associated with preterm birth, and we have chosen 11 common clinical predictors and 4 genetic risk predictors as an example for visualization.

Clinical predictors such as extremes of age have shown to have an estimated odds of 1.8 [18], and low weight gain is strongly associated with PTB with an adjusted odds ratio of 9.8, while the odds are doubled for women with a high weight gain [19]. Smoking 10 cigarettes a day during pregnancy will also triple the risk of PTB [20]. Moreover, mothers who were born preterm themselves have an estimated odds ratio of 1.18 for delivering preterm compared to mothers born at term [21].

Interleukin-1β ($IL1\beta$) and Interleukin-6 ($IL6$), which encode pro-inflammatory cytokines that affect gestational tissues, appear to be the most consistent genes that have been associated with PTB [22, 23]. Other genes such as Transforming Growth Factor β (TGFβ), which is also linked with inflammation, and Factor V Leiden (F5), which is linked with thrombosis, have also been studied [19, 24, 25].

The Chernoff faces shown in Fig.1 are based on standardized data, as the predictors are of different units and scales. The face characteristics and corresponding clinical and SNP predictors are shown in Table 1. Each face represents a patient, with the patient study ID shown at the top. Patients who delivered preterm are highlighted with a yellow face.

With all the predictors plotted into faces, it is relatively easy to group patients with similar characteristics. For instance, smiling faces indicate patients who are still using Marijuana at 15 weeks of gestation, and faces with closed-eyes indicate patients who had a low Socioeconomic Index and are depressed. Similarly, faces with a wide-open mouth indicate patients who consumed alcohol and are obese.

Patients who had uncomplicated pregnancies are expected to have a longer Cervical length compared to PTBs, which are reflected by a wider Chernoff face. From Figure 1, it appears that most PTB cases appears to have 'thin' or 'small' faces, reflecting a short cervical length and younger age, with the exception of case 795 where the face appears to be relatively tall, yet in

Characteristics	Predictors
Face height	Age
Face width	Cervical length
Face structure	Gravidity
Mouth height	Alcohol (1st trimester)
Mouth width	Body Mass Index
Smiling	Marijuana (1st visit)
Eye height	Social Economic Index
Eye width	Depression
Hair height	Cigarettes (1st visit)
Hair width	Anxiety index
Hair style	Born preterm
Nose height	Interleukin-6
Nose width	Factor V Leiden
Ear height	Transforming Growth Factor β
Ear width	Interleukin-1β

Table 1: Facial characteristics and corresponding predictors

fact, she is the oldest patient.

More distinct facial characteristics between Uncomplicated pregnancies and PTBs can be seen by the absence of ears and high hair style. The faces of patients who delivered PTB have either a tall hair style or no ears, which correspond to whether the mother was herself born preterm, and having a particular genotype in TGFβ and IL1β genes.

5. Conclusions

As shown, multivariate visualization of SNPs and clinical predictors can be achieved simultaneously through Chernoff faces. Patients with similar characteristics can be easily identified through similar facial characteristics. This feature is particularly useful in developing grouping or clustering methods for disease through combinations of clinical and SNP predictors.

References

[1] J. Perkel, Snp genotyping: six technologies that keyed a revolution., Nature methods 5 (2008) 447–453.

[2] A. Johnson, Single-nucleotide polymorphism bioinformatics., Circulation: Cardiovascular Genetics 2 (5) (2009) 530–536.

[3] J. Gibbs, A. Singleton, Application of genome-wide single nucleotide polymorphism typing: Simple association and beyond., PLoS Genet 2 (10) (2006) e150.

[4] I. Gray, D. A. Campbell, N. K. Spurr, Single nucleotide polymorphisms as tools in human genetics., Human Molecular Genetics 9 (16) (2000) 2403–2408.

[5] N. Shah, M. V. Teplitsky, S. Minovitsky, L. A. Pennacchio, P. Hugenholtz, B. Hamann, I. L. Dubchak, Snp-vista: An interactive snp visualization tool., BMC Bioinformatics 6 (1) (2005) 292.

[6] F. Sørensen, C. Andersen, C. Wiuf, Snptools: a software tool for visualization and analysis of microarray data., Bioinformatics 23 (12) (2007) 1550–1552.

[7] O. Martin, A. Valsesia, A. Telenti, I. Xenarios, B. Stevenson, Associationviewer: a scalable and integrated software tool for visualization of large-scale variation data in genomic context., Bioinformatics 25 (5) (2009) 662–663.

[8] K. Bhasi, L. Zhang, D. Brazeau, A. Zhang, M. Ramanathan, Vizstruct for visualization of genome-wide snp analyses., Bioinformatics 22 (13) (2006) 1569–1576.

[9] H. Chernoff, The use of faces to represent points in k-dimensional space graphically, Journal of American Statistical Association 68 (342) (1973) 361–368.

[10] The scope pregnancy research study.
URL `http://www.scopestudy.net/`

[11] K. Hanretty, Obstetrics Illustrated, 7th Edition, Churchill Livingstone, 2009.

[12] S. Pfeifer, NMS Obstetrics and Gynecology: National Medical Series for Independent Study, 6th Edition, Lippincott Williams & Wilkins, 2007.

[13] M. Kramer, K. Demissie, H. Yang, R. W. Platt, R. Sauvé, R. Liston, The contribution of mild and moderate preterm birth to infant mortality. fetal and infant health study group of the canadian perinatal surveillance system., Jama 284 (7) (2000) 843–9.

[14] K. Mikkola, N. Ritari, V. Tommiska, T. Salokopi, L. Lehtonen, O. Tammela, L. Pääkkönen, P. Olsen, M. Korkman, V. Fellman, Neurodevelopmental outcome at 5 years of age of a national cohort of extremely low birth weight infants who were born in 1996-1997., Pediatrics 116 (6) (2005) 1391–400.

[15] A. Den Ouden, J. H. Kok, P. Verkerk, R. Brand, S. P. Verloove-Vanhorick, The relation between neonatal thyroxine levels and neurodevelopmental outcome at age 5 and 9 years in a national cohort of very preterm and/or very low birth weight infants., Pediatr Res 39 (1) (1996) 142–5.

[16] H. Honest, C. A. Forbes, K. H. Durée, G. Norman, S. B. Duffy, A. Tsourapas, T. E. Roberts, P. M. Barton, S. M. Jowett, C. J. Hyde, K. S. Khan, Screening to prevent spontaneous preterm birth: systematic reviews of accuracy and effectiveness literature with economic modelling., Health Technol Assess 13 (43) (2009) 1–627.

[17] R. E. Behrman, A. S. Butler (Eds.), Committee on Understanding Premature Birth and Assuring Healthy Outcomes, Preterm Birth: Causes, Consequences, and Prevention, Washington DC: National Academies Press, 2007.

[18] J. A. Martius, T. Stecka, M. K. Oehlerb, K. Wulfa, Risk factors associated with preterm (< 37+0 weeks) and early preterm birth (< 32+0 weeks): univariate and multivariate analysis of 106 345 singleton births from the 1994 statewide perinatal survey of bavaria., Eur J Obstet Gynecol Reprod Biol 80 (2) (1998) 183–9.

[19] D. Murphy, Epidemiology and environmental factors in preterm labour., Best Practice & Research Clinical Obstetrics & Gynaecology 21 (5) (2007) 773–789.

[20] A. Braillon, S. Bewley, The enigma of spontaneous preterm birth., N Engl J Med 362 (21) (2010) 2032, author reply 2033-4.

[21] T. F. Porter, A. M. Fraser, C. Y. Huntera, R. H. Ward, M. W. Varner, The risk of preterm birth across generations., Obstet Gynecol 90 (1) (1997) 63–7.

[22] R. L. Goldenberg, J. F. Culhane, J. D. Iams, R. Romero, Epidemiology and causes of preterm birth., Lancet 371 (9606) (2008) 75–84.

[23] S. A. Engel, H. C. Erichsen, D. A. Savitz, J. Thorp, S. J. Chanock, A. F. Olshan, Risk of spontaneous preterm birth is associated with common proinflammatory cytokine polymorphisms., Epidemiology 16 (4) (2005) 469–77.

[24] N. Chegini, J. Davis, P. Duff, C. Rosa, Differential expression of transforming growth factor-beta 1 and transforming growth factor-beta receptors in myometrium of women with failed induction of labor, no labor, and preterm labor., J Soc Gynecol Investig 6 (5) (1999) 258–63.

[25] R. Mattar, E. de Souza, S. Daher, Preterm delivery and cytokine gene polymorphisms., J Reprod Med 51 (4) (2006) 317–20.

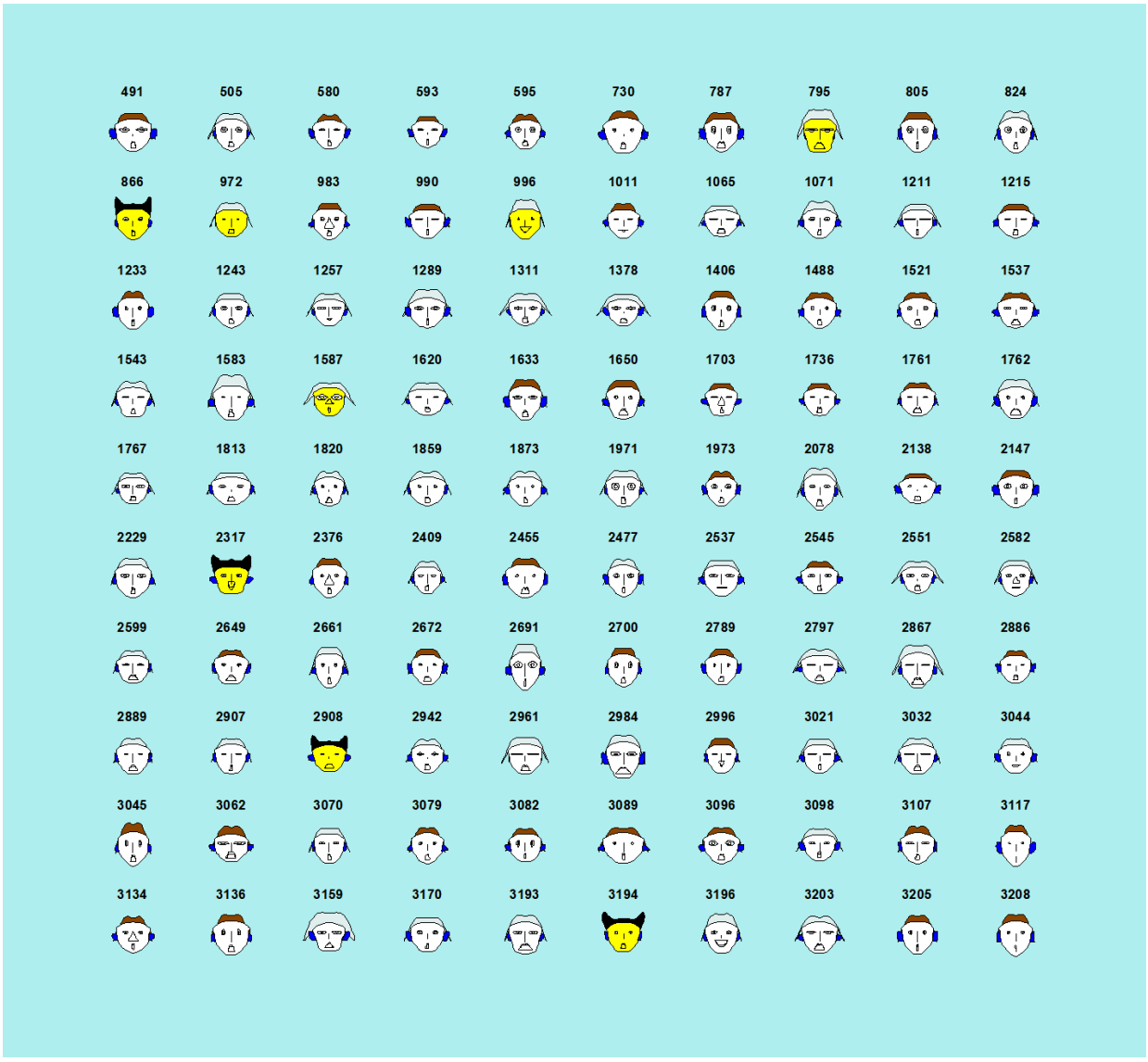

Figure 1: Chernoff faces displaying 11 clinical and 4 genetic predictors for PTB (with PTB cases highlighted)