Προσχέδιο Διδακτορικής Διατριβής

Τίτλος: Model Selection in Generalized ROC studies.

Χρονική Διάρκεια Ερευνητικού Έργου : Από σαράντα οκτώ (48) έως εβδομήντα δύο (72) μήνες.

1 έτος, Επισκόπηση βιβλιογραφίας και αναλυτική διερεύνηση νέων τεχνικών στις ROC καμπύλες.

2 έτος, Ανάπτυξη νέων μεθοδολογιών.

3 έτος, Σύγκριση των προτεινομένων μεθοδολογιών με τις ήδη υπάρχουσες. (Simulation studies και εφαρμογές σε πραγματικά δεδομένα. Έναρξη συγγραφής επιστημονικών άρθρων).

4 έτος, Δημοσιεύσεις και περάτωση διδακτορικής διατριβής.

Αντικείμενο και Στόχοι του Προτεινόμενου Έργου

Ο ρόλος της βιοστατιστικής καθώς και η σημασία της στις κλινικές μελέτες εμφανίστηκε τον 17^{ο} αιώνα περίπου, κάνοντας την αναγκαία όσο περνούν τα χρόνια, τόσο για την διόρθωση σφαλμάτων που τυχόν υπήρχαν από τους ερευνητές, όσο και για την διαγνωστική ιατρική. Οι τελευταίες εξελίξεις σε παγκόσμιο επίπεδο μετά από την εμφάνιση και του νέου κορωνοϊού (COVID-19) έφεραν στην επιφάνεια για μία ακόμη φορά την ανάγκη της χρήσης βιοστατιστικής για την ανάπτυξη νέων στατιστικών μεθόδων για την άμεση αλλά και αποτελεσματική αντιμετώπιση ασθενειών που προσβάλλουν το γενικό πληθυσμό. Ένα από τα ζητήματα που έχουν απασχολήσει την επιστημονική κοινότητα αφορά τους διαγνωστικούς ελέγχους. Οι διαγνωστικοί έλεγχοι χρησιμοποιούνται για τον έλεγχο, την ανίχνευση αλλά και την παρακολούθηση ασθενειών φανερώνοντας έτσι τον κύριο ρόλο που έχουν στις αποφάσεις οι οποίες λαμβάνονται για την υγεία. Πολλοί ερευνητές υποστηρίζουν ότι η έγκαιρη διάγνωση μπορεί να επιφέρει αντιμετώπιση ή και θεραπεία οποιασδήποτε νόσου με μεγαλύτερη επιτυχία μέσα από ελέγχους που βοηθούν στην κατανόηση του μηχανισμού και της φύσης μιας νόσου. Επίσης η συλλογή αυτών των πληροφοριών μπορεί να χρησιμοποιηθεί για τη βελτίωση της υγείας του γενικού πληθυσμού. Με αυτόν τον τρόπο γίνεται εμφανής η ανάγκη για εύρεση μεθόδων που θα εξασφαλίσουν την αποτελεσματικότητα των ελέγχων και της έρευνας, καθώς και την θέσπιση ενός θεωρητικού πλαισίου για την σωστή ερμηνεία των αποτελεσμάτων.

Αυτή η ερευνητική πρόταση εστιάζει την προσοχή της στην αποτελεσματικότητα της διαγνωστικής ακρίβειας. Το κατεξοχήν εργαλείο που χρησιμοποιείται στην αξιολόγηση διαγνωστικών ελέγχων και που καταλήγει στην ταξινόμηση ατόμων σε ένα από τα στάδια μιας ασθένειας είναι οι Receiver Operating Characteristic (ROC) καμπύλες. Αυτές αρχικά χρησιμοποιήθηκαν στην ανίχνευση σήματος (signal detection) (Green And Swets 1966). Ο Lusted (1971) εισήγαγε στην διαγνωστική ιατρική μια

μέθοδό περιγραφής της εγγενούς ακρίβειας ενός ελέγχου, η οποία υπερνικά το πρόβλημα λαμβάνοντας υπόψη όλα τα πιθανά σημεία απόφασης. Για περεταίρω πληροφορίες μπορεί κανείς να διαβάσει τα άρθρα των Zhou et al. (2002) and Pepe (2003). Μαθηματικά η κλασσική ROC καμπύλη περιέχει όλη την πληροφορία που χρειάζεται κανείς για την σύγκριση 2 κατανομών. Το εμβαδόν κάτω από την καμπύλη (AUC) είναι το πιο ευρέως διαδεδομένο περιληπτικό μέτρο που χρησιμοποιείται στους διαγνωστικούς ελέγχους και είναι άμεσα συνδεδεμένο με τον έλεγχο Wilcoxon Rank Sum. Γενικεύσεις των ROC καμπυλών παρουσιάζουν ιδιαίτερο ενδιαφέρον τα τελευταία χρόνια. Μία γενίκευση αφορά στην ταξινόμηση σε πάνω από δύο κατηγορίες, για παράδειγμα (υγιείς, στάδιο 1, στάδιο 2, στάδιο 3). Μία περαιτέρω γενίκευση αφορά την αξιολόγηση διαγνωστικών ελέγχων με δύο ή παραπάνω κατηγορίες και λαμβάνει υπόψη της τα κόστη λάθος ταξινόμησης (misclassification costs). Κατά την τελευταία εικοσαετία έχουν αναπτυχθεί μέθοδοι που καταπιάνονται με το γενικότερο πρόβλημα των πολλαπλών κατηγοριών (Nakas 2004). Επίσης συνεχίζεται η ανάπτυξη νέων μεθόδων που συνδέουν την ακρίβεια ενός διαγνωστικού ελέγχου με συμμεταβλητές.

Οι στόχοι αυτού του έργου είναι οι εξής:

(1) Ανάπτυξη παραμετρικών και ημι-παραμετρικών μοντέλων για γενικές μελέτες ROC (με πολλαπλές κατηγορίες ασθένειας) υπό την παρουσία συμμεταβλητών.

(2) Εκτίμηση κανόνων απόφασης όπου θα λαμβάνεται υπόψη το κόστος λάθος ταξινόμησης.

(3) Εξέταση συνεχών και διακριτών κατανομών όσον αφορά την κατανομή του marker.

(4) Επιλογή Μοντέλου.

Βιβλιογραφία

Banerjee, P., Garai, B., Mallick, H., Chouwdhury, S., and Chtterjee, S. (2018). A Note on the Adaptive LASSO for Zero-Inflated Poisson Regression. *Journal of Probability and Statistics.* https ://doi.org/10.1155/2018/28341 83.

Bantis, L.E., Tsimikas, J.V., and Georgiou, S.D. (2011). Survival estimation through the cumulative hazard with constrained natural cubic splines. *Lifetime Data Analysis* **18(3)**, 364-396.

Bantis, L.E., Tsimikas, J.V., and Georgiou, S.D. (2011). Smooth ROC curves and surfaces for markers subject to a limit of detection using monotone natural cubic splines. *Biometrical Journal*.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical Report, University of California, Berkeley.*

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52:3**, 345-370.

Cai, T., Pepe, M. S., Lumley, T., Zheng, Y., and Jenny, N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics* **7,** 187-197.

Claeskens, G., Hjort, N. L. (2008). Model Selection and Model Averaging. Cambridge University Press.

Efron, B. (1967). The two sample problem with censored data. *In Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probability, New York: Prentice-Hall* **4**, 831-853.

Fleming, T. R., and Harrington, D. P. (1990). Counting processes and survival analysis. *Wiley Series in Probability and Statistics*.

Fritsch F. N., and Carlson, R. E. (1980). Monotone piecewise cubic interpolation *SIAM Journal of Numerical Analysis* **2**, 238-246.

Gentleman, R., and Crowley, J. (1991). Graphical methods for censored data *Journal of the American Statistical Association* **86**, 678-683.

Gonen, M., and Heller, G. (2010). Lehmann family of ROC curves. *Medical Decision Making* **30**, 509-517.

Green, D.M., and Swets, J.A. (1966). *Signal detection theory and psychophysics* Wiley, New York.

Heagerty, P.J., Lumley, T., and Pepe, M.S. (2000). Time-Dependent ROC Curves for censored survival data and a diagnostic Marker. *Biometrics* **4**, 337-344.

Herndon, E. J. II, and Harrell F. E. Jr. (1995). The restricted cubic spline as baseline hazard in the proportional hazards model with step function time-dependent covariables. *Statistics in Medicine* **14**, 2119-2129.

Heyde, C.C. (1997). Quasi-Likelihood and its application: A general approach to optimal parameter estimation. Springer-Verlag New York.

Hornung, W. R., and Reed, D.L. (1990). Estimation of average concentration in the presence of nondetectable values *Applied Occupational and Environmental Hygiene* **5**, 46-51.

Hughes, M.D. (2000). Analysis and design issues for studies using censored biomarker measurements with an example of viral load measurements in HIV clinical trials. *Statistics in Medicine* **19**, 3171-3191.

Jiang, Y., Metz, E.C., and Nishikawa, R.M. (1996). A Receiver Operating Characteristic partial area index for highly sensitive diagnostic tests *Radiology* **201**, 745-750.

Kardaun, O. (1983). Statistical analysis of male larynx cancer patients. A Case Study. *Statistical Nederlandica* **37**, 103-126.

Klein, J.P., and Moeschberger, M.L. (2003). Survival Analysis, Techniques for censored and truncated data. *Springer Verlag*.

Kooperberg, C., Stone, J. C., and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association* **90**, 78-94.

Krzanowski, W. J., and Hand, D. J. (2009). ROC curves for continuous data. London: Chapman and Hall.

Leisenring, W., Pepe, M.S., and Longton, G.L. (1997). A marginal regression modeling framework for evaluating medical diagnostic tests. *Statistics in Medicine* **16**, 1263-1281.

Lehmann, E.L. (1953). The power of rank tests. *Ann. Math. Stat.* **24**, 23-43.

Lu, H., Xu, Y., Ye, M,. Yan, K., Gao, Z., and Jin, Q. (2019). Learning misclassification costs for imbalanced classification on gene expression data. *BMC Bioinformatics* **20**, 681. https://doi.org/10.1186/s12859-019-3255-x

Lusted, L. B. (1971). Signal detectability and medical decision making. *Science* **171**, 1217-1219.

Ma, H., Halabi, S., and Liu, A. (2019). On the Use of Min-Max Combination of Biomarkers to Maximize the Partial Area under the ROC Curve. *Journal of Probability and Statistics.* https://doi.org/10.1155/2019/8953530

Nakas, C.T., and Yannoutsos, C.T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine* **23**, 3437-3449.

Naylor, S. (2005). Overview of biomarkers in disease, drug discovery and development *Drug Discovery World Spring* 21-30.

Pepe, M.S. (2003). *The statistical evaluation of medical tests for classification and prediction.* Oxford University Press.

Perkins, N.J., Schisterman, E.F., and Vexler, A. (2006). A Receiver Operating Characteristic Curve inference from a sample with a limit of detection. *American Journal of Epidemiology* **165(3),** 325-333.

Prentice, R.L. (1982). Covariate measurement errors and parameter estimation in failure time regression models. *Biometrika* **69,** 331-342.

Stablein, D. M., and Koutrouvelis, I.A A (1985). A two-sample test sensitive to crossing hazards in uncensored and singly censored Data. *Biometrics* **41**, 643-652.

Stacy, E.W. (1962). A Generalization of the Gamma Distribution. *Annals of Mathematical Statistics* **33(3)**, 1187-1192.

Steelman, C.D., Gbur, E.E., Tolley, G., and Brown, A.H.Jr (1993). Variation in population density of the face fly, Musca autumnalis De Geer, Among Selected Breeds of Beef Cattle. *J. Agric. Entomol.* **10(2)**, 97-106.

Stone, C. J. (1990). Large sample inference for log-Spline models. *The Annals of Statistics* **18**, 717-741.

Tsiatis, A., and Davidian, M. (2004). Joint modeling of longitudinal and time-to-event data: An overview. Statistica Sinica **14**, 809-834.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated Data. *Journal of the Royal Statistical Society (Series B)* **38**, 290-295.

Tusch G. Evaluation of partial classification algorithms using ROC curves. Medinfo. 1995;8 Pt 2:904-8. PMID: 8591580.

Wang, Z., and Chang, Y. C. I. (2011). Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics* **12(2)**, 369-385.

Wassertheil-Smoller, S. (1990). Biostatistics and epidemiology. Springer-Verlag New York.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton Method. *Biometrika* **61**, 439-447.

Zhang Y., Hua L., and Huang J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics* **37**, 338-354.

Zhou, X.H., Obuchowski, N.A., and McClish, D.K. (2002). *Statistical methods in diagnostic medicine* Wiley, New York.

Zou, K.H., Liu, A., Bandos, I., Ohno-Machado, L., and Rockette, H.E. (2011). Statistical Evaluation Of Diagnostic Performance: Topics in ROC Analysis. Boca Raton, FL: Chapman & Hall/CRC Biostatistics Series.