Using Decision Trees to Find Patterns in an Ophthalmology Dataset

Susan P. Imberman Ph.D [1] ., Irene Ludwig, M.D [2] ., Sarah Zelikovitz, Ph.D. [3]

[1,3]College of Staten Island, Graduate Center, City University of New York
2800 Victory Blvd.
Staten Island, New York 10314
[1]susan.imberman@csi.cuny.edu, [2]ihludwig@aol.com , [3]sarah.zelikovitz@csi.cuny.edu

Abstract

We present research in decision tree analysis that studies a data set and finds new patterns that were not obvious using statistical methods. Our method is applied to a database of accommodative esotropic patients. Accommodative esotropia is an eye disease that when left untreated leads to blindness. Patients whose muscles deteriorate often need corrective surgery, since less invasive methods of treatment tend to fail in these patients. Using a learn and prune methodology, decision tree analysis of 354 accommodative esotropic patients led to the discovery of two conjunctive variables that predicted deterioration in the initial year of treatment better than what was previously determined using standard statistical methods.

Introduction

Traditionally, the analysis of medical clinical studies has been done using standard statistical tests. Statistical analysis requires that the researcher know the set or subset of data variables that need to be analyzed. Alternatively, data mining can lead to the discovery of variable relationships in data that may not have been realized or previously identified [1]. KDD, Knowledge Discovery in Databases, is a multi-step process that analyzes data with respect to the discovery of patterns rather than the rejection of hypotheses. Data mining is one step in this process.

Esotropia (crossed eyes) is an eye disorder in children that can result in blurred or double vision [2, 3]. In order to accommodate for these vision deficiencies the child tends to suppress the vision in one eye, possibly leading to amblyopia (blindness). Esotropia is treated by using corrective lenses, miotic agents (eye drops), eye patching, or surgery. Surgery is used when less invasive methods are ineffective. It would be extremely useful for physicians to know which children will eventually require surgery, but this is difficult to predict. A previous clinical study identified several risk factors for surgery [4]. A comprehensive clinical study with 19,000 records describing 1,307 patients with 40 different variables tracked was subsequently conducted [5].

The work described in this paper applies machine learning techniques, specifically decision trees, to this large clinical dataset. This paper describes the use of decision trees to find patterns in this data, and helps give physicians insight into the the data that was previously not realized using traditional methods.

Methodology

The data mining process begins with a cycle of data cleaning, analysis and consultation with domain experts. Following the data set-up, we apply our machine learning algorithm, which in this paper was decision trees. Based on the output of our learning algorithm, a predictor variable is chosen. Some variables are calculated variables, i.e. combinations of multiple attributes. If the predictor is a calculated variable, we reexpress this predictor using various combinations of the calculated variable and its component variables, once again using feedback and consultation from domain experts. These new predictors are then compared to the parent predictor so that the best level of granularity can be chosen. This process is iterative, and the newly found predictor can then be used to prune the data, so that the process can begin again. Data was pruned either by removing patients meeting some criteria, or by running the learning algorithm on a smaller attribute set. As a result, the process yields insight into subgroups within the data, and rules that hypothesize over these subgroups.

The study dataset has information that was obtained during patient visits to an ophthalmologist's office. It tracks 1307 patients, with an average of 14 visits per patient. The patient with the most visits had 52 visits, with the fewest number of visits equal to 2. By studying multiple lines of data for one specific patient, a physician can follow the patient's progress, or deterioration over time. The data set consisted of 54 eye related attributes, including some measured by the physician during visits, and others calculated based on a combination of the measured features.

Three hundred and fifty four patients exhibiting accommodative esotropia (cross-eyes) were selected from the study dataset. The average number of records per patient in the accommodative esotropia subset was 14,

with 5,073 records total. We summarized the initial year's worth of data for each patient to obtain one record per patient based upon all visits during that year.

The J4.8 decision tree algorithm was chosen to do initial analysis of the first year's data [6]. We preferred to use decision trees because of their transparency and easy rule-creation properties that can provide data that is understandable and useable by physicians. Since we interfaced with an ophthalmologist throughout this process, it was of utmost importance to us that the output of the algorithm we chose was understandable to the experts. *Why* a child is predicted to need surgery in the future is as important to the ophthalmologist, and the child's other care providers, as the fact that she may or may not need surgery.

Results and Discussion

In consultation with a domain expert, one data attribute that correlated well with deterioration is the *AC/A* ratio, a calculated field that describes the ability of the eyes to focus. However, domain experts have a biased view of the data. By using decision trees we are able to take an unbiased look at the entire dataset, where in addition to calculated variables, we look at their component parts in the same analysis. Because of this approach we are able to get new combinations of variables, in addition to other insights into the data.

We were **not** looking to build a good classifier, but to find patterns that indicated deterioration. Our domain expert indicated that there were most probably patient subgroups in the data where, even though the disease presented similarly, the underlying process differed. We were hoping to identify features of some of these subgroups.

Initial analysis with the full 354 patient dataset and J4.8 produced a tree with 73 nodes, of which 43 were leaves. Interestingly, a distance vision measurement called *distCC1* (measurement of distance vision using corrective lenses) was picked over *AC/A* in classifying the root node. The root node of the tree was split at a value of distance greater than 8, indicating that those patients with *distCC1* > 8 might have additional features that indicate deterioration.

We pruned the attribute list such that *distCC1* was NOT used to build the tree. The resulting tree had 42 nodes of which 23 were leaves. Here, another distance vision measurement called *distCC2* labeled the root node. The difference between *distCC2* and *distCC1* is that when measuring *distCC2*, the eyes are fatigued by continually repeating the distance exam. The values at which *distCC2* split was, like *distCC1*, a value of 8.

Using the insights obtained from this analysis and feedback from the domain expert, it was found that patients exhibiting both a *distCC1* with values between 6 and 8 inclusive, and a *distSC* (measurement of distance vision without corrective lenses) value of less than or equal to 3, predicted deterioration with a sensitivity of 94% and a specificity of 42% which was better than that obtained from the *AC/A* ratio on this subset, having a 89% sensitivity and a 37% specificity. According to the domain expert, this meant that we were able to identify a subgroup of patients with constant or intermittent esotropia, or poor control over the eye muscles, who deteriorate more readily than patients with esophoria, or good control over the eye muscles.

Summary and Conclusions

Decision trees were used to find patterns leading to deterioration in a clinical ophthalmology dataset. We continually pruned the dataset and reapplied our learning algorithm according to a prescribed methodology. This analysis led to the finding of two variables that conjunctively predicted deterioration with better sensitivity and specificity than that previously determined using observation and standard statistical techniques. We were able to identify a subgroup of patients, in the early stages of treatment, who tend to deteriorate. We plan to continue analyzing the data using other techniques such as cluster analysis and association rules.

This work was supported by PSC-CUNY grant..62323-00-40.

References

[1] Lavarc, N, Keravnou, E. T., Zupan, B. , 1997. *Intelligent Data Analysis in Medicine and Pharmacology,* Kluwer Academic Publishers.

[2] Vaughan, Daniel, Asbury, Taylor, 1986. *General Ophthalmology,* Lange Medical Publications

[3] Von Noorden, Gunter K. 1977. *Atlas of Strabismus*, The C. V. Mosby Company

[4] Ludwig, I.H., Imberman, S. P, Thompson, H. , Parks, M. M., *Long Term Study of Accommodative Esotropia, December 2005,* Journal of the American Association for Pediatric Ophthalmology and Strabismus, Vol 9, Issue 6, pgs. 522-526

[5] Ludwig I. H, Parks M. M, Getson P. R, Kammerman L. A.. *Rate of deterioration in accommodative esotropia correlated to the AC/A relationship*. Journal Pediatric Ophthalmology Strabismus 1988; 25:8-12.

[6] Hall, M., l Frank, E, Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H (2009); The WEKA Data Mining Software: An Update, SIGKDD Explorations, Volume 11, Issue1