Detection of Sentinel Predictor-Class Associations With XCS: A Sensitivity Analysis

John H. Holmes
University of Pennsylvania
School of Medicine
Philadelphia, PA 19104
+1-215-898-4833

jholmes@cceb.med.upenn.edu

ABSTRACT

Knowledge discovery in databases has traditionally focused on classification, prediction, or in the case of unsupervised discovery, clusters and class definitions. Equally important, however, is the discovery of individual predictors along a continuum of some metric that indicates their association with a particular class. This paper reports on the use of an XCS learning classifier system for this purpose. Conducted over a range of odds ratios for a fixed variable in synthetic data, it was found that XCS discovers rules that contain metric information about specific predictors and their relationship to a given class. In addition, EpiXCS performs qualitatively similarly to See5, and both methods are comparable to logistic regression.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning– *concept learning, induction, parameter learning*

General Terms

Algorithms, Experimentation, Performance, Reliability.

Keywords

Learning classifier systems, variable selection, statistical analysis, statistical computing

1. INTRODUCTION

A number of study designs exist for the collection and analysis of epidemiologic surveillance data. These range from static, one-time observational (prevalence) studies of an existing population or sample, to ongoing observation of a given population or sample. The latter, commonly called a *cohort study*, is of particular interest, in that it provides the ability to investigate the incidence of outcomes of interest over time. Because of this added informational dimension, it is possible to create more robust models of causation that as would be the case in simple observational studies. However, cohort studies are expensive to maintain, and especially with regard to rare outcomes, may require many years of observation. The alternative to the cohort design is the *case-control study*, in which subjects, who possess an outcome of interest, are identified as *cases*. An equal or greater number of subjects without the outcome are included in the study as *controls*. The controls are typically selected at random from a similar population as the cases, and may be matched on some set of variables, such as age, sex, or race. A case-control study may be performed within the context of a larger cohort. The advantage of the case-control study is that it is particularly efficient for rare outcomes which would take too long to complete in the context of a cohort study.

One of the essential analytic parameters in the case-control study is the odds ratio (OR). The OR approximates the relative risk of an outcome associated with a given exposure [3]. For dichotomous exposure and outcome variables, it is calculated as a cross-product ratio, as shown in Figure 1.

Exposure	Class	
	Cases	Controls
Exposed	A	B
Not Exposed	C	D

Figure 1. A 2x2 table showing exposure by case-control status. "Exposed" indicates whether or not a given variable of interest was positive or not. "Cases" are those with an outcome of interest, while "Controls" are those without. The crude (unadjusted) OR is calculated as AD/BC.

Odds ratios of less than 1 indicate a *protective effect* of the predictor of interest on clinical outcome; while ORs greater than 1 indicate a positive association of the exposure with the outcome. If the OR is 1.0, the predictor has no association, positive or protective, with the outcome. The OR described in the above figure is commonly referred to as the *crude odds ratio*, because it has not been adjusted for other variables that may be important confounders or effect modifiers. While the crude OR is useful, the adjusted OR, typically derived by means of a logistic regression model, is of more importance in ascertaining the true effect of a given variable on an outcome. However, logistic models are not without some degree of weakness: small sample sizes, missing data, and model complexity that may cause failure of the model to converge are some of the reasons why important variables might not be identified as statistically significant. This is especially the case in the early phases of epidemiologic surveillance, where the number of outcome associations may be small and apparently random. The discovery of features (variables as well as specific values of variables) that act as

Gecco'05, June 25-29, 2005, Washington, DC, USA.

sentinels in alerting investigators to potential relationships is of great importance to epidemiologic investigators.

This paper describes an ongoing investigation into the ability of an XCS-type learning classifier system to identify such feature-outcome associations in a simulated case-control study under strict experimental conditions.

2. METHODS

2.1 Data

2.1.1 Baseline Data

A small case-control study was simulated by creating a dataset consisting of 10 real-valued predictor variables and one dichotomous outcome variable. The value of each predictor was randomly generated using a bounded normal distribution (range 0-10). The outcome variable was also randomly generated, using a binomial distribution, constrained such that the dataset contained 100 records, with equivalent class frequency to yield 50 "cases" and 50 "controls." Throughout this paper, the positive class is referred to as *Cases* and the negative class, *Controls*. The random data generation procedures ensured that no associations existed between any predictors and the class variable, verified by one-sample t-test ($p>0.05$). In addition, the crude and adjusted ORs for each variable were calculated and found to be approximately 1.0; their confidence intervals were not statistically significant. The baseline dataset is described below in Table 1.

Table 1. The baseline dataset used for this investigation.

Variable	Mean (Standard Deviation)	Minimum	Maximum
1	4.99 (2.85)	0	9.7
2	4.57 (2.91)	0.2	10.0
3	4.79 (2.80)	0.1	9.8
4	5.23 (2.82)	0.2	10.0
5	5.11 (3.01)	0.1	10.0
6	4.77 (2.79)	0.1	9.9
7	5.28 (2.85)	0.2	9.8
8	5.18 (2.73)	0.3	9.9
9	5.75 (2.96)	0	10.0
10	4.73 (2.90)	0.4	10.0
Class	50 "cases" and 50 "controls"		

2.1.2 Incremental Data

The goal of this investigation was to evaluate the ability of XCS to discover rules that indicate an increased OR over a range of values. The baseline dataset was altered to increase the crude OR for a single variable (Attribute 2) from 1.0 to 4.0. The goal was to associate values greater than or equal to 2.0 with Cases at a gradually increasing rate. This was accomplished by randomly selecting a Case record with a value greater than 2.0 for Attribute 2, and changing the value of Attribute 2 for this record to a random value between 0 and 2.0. Each iteration was successively written to a different dataset for evaluation. Thus, each dataset contained the alterations of the one created before it, in addition to the alteration at that iteration. A total of six such datasets were created, with crude ORs of 1.56, 2.06, 2.45, 2.90, 3.41, and 4.0.

2.1.3 Experimental procedure

This investigation focused on rule discovery occurring during the training phase in a supervised learning environment. As a result, no training-testing set pairs were created; however, 10-fold cross-validation was used to ensure random and complete exposure to each record in the datasets. EpiXCS [2] was used as the learning classifier system for the experiments. Parameters were set as per [1]. Decision tree and rule induction was performed with See5 [4]. Crude and adjusted ORs were determined by the *epitable* and *logistic* procedures in STATA, version 8.0 [5].

Each dataset was trained in EpiXCS over five runs. The rulesets described here represent the conflation of these sets into a "meta-set" that was created by ranking the rules by their predictive value. The rulesets derived by EpiXCS were evaluated using the visualization tool provided in this software. Of particular interest was the degree of "scatter" over the value ranges for each variable: highly scattered values indicate randomness and lack of specificity. However, highly bounded values represent possible associations with the class of interest. The rule evaluation focused primarily on Cases, but also Controls, in an effort to identify rule patterns that discriminated between them.

The rulesets derived by See5 were likewise examined qualitatively over the range of ORs, to identify the emergence of Attribute 2 as a possible predictor of the class. The ORs calculated by the logistic regressions were examined for magnitude and statistical significance.

3. RESULTS

3.1 Rule discovery

3.1.1 EpiXCS

The rules discovered by EpiXCS on the baseline data for the cases are shown in Figure 2. The rule visualization tool shown in this figure is read as follows. The bottom pane lists the rules discovered at the end of training, after they have been collected from each of the five runs and sorted by their predictive value. The predictive value is either positive or negative, depending on the class that is being displayed. In Figure 2, the ruleset for the Cases are shown; thus, the rules are sorted by their positive predictive value. The text of each rule is color coded to facilitate its visualization on the graph that is shown in the top pane. The horizontal axis of this graph represents the predictors in the dataset. The vertical axis represents the possible value range for the predictors. In this dataset, all predictors ranged from 0 to 10; however, in data where the ranges are variable, the minimum and maximum values taken over all variables are used for this scale. The colored bars indicate the ranges of the values of the variables participating in a given rule. Again, a given bar is mapped by color to its rule in the lower pane. Single or multiple rules may be visualized at once, as is the case in Figure 2.

As expected, the rules in the baseline data are characterized by a random pattern, in that no one variable or set of variables seems to be associated with the class. A similar rule pattern was found on the rulesets at the lower ORs. Figure 3 shows the inverse of what would be expected for the value of Attribute 2 as associated with the class: the best-performing rules for this class indicate that the value of Attribute 2 should be greater than 2.

However, the expected value for Attribute 2 emerges with increasing OR, as is shown in Figure 4. In this figure, obtained at

OR=2.90, the value for Attribute 2 ranges from 0 to 5. The emergence of the expected value is complete at OR=3.41, as shown in Figure 5. Figure 6 shows the ruleset derived for the Controls at this OR. In comparing the shaded portions of each graph for Attribute 2 (shown as Attribute 2), one can clearly see the effect of dichotomization of Attribute 2 at values <=2. In Figure 5, the red bar does not exceed a value of 2, while in Figure 6, the lavender bar for the same variable covers approximately the complementary values (about 2.5 to 10).

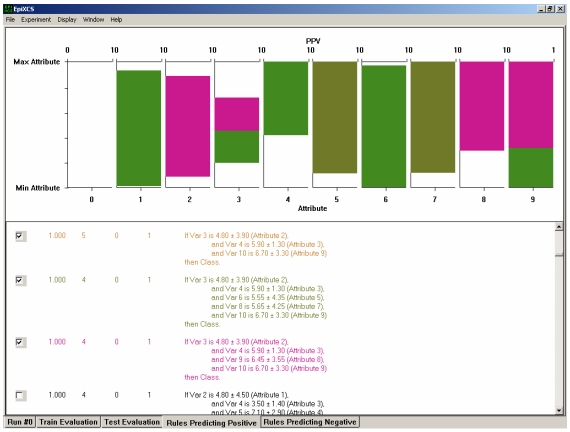

Figure 2. Ruleset derived by EpiXCS on the baseline data. The display is described in the text.

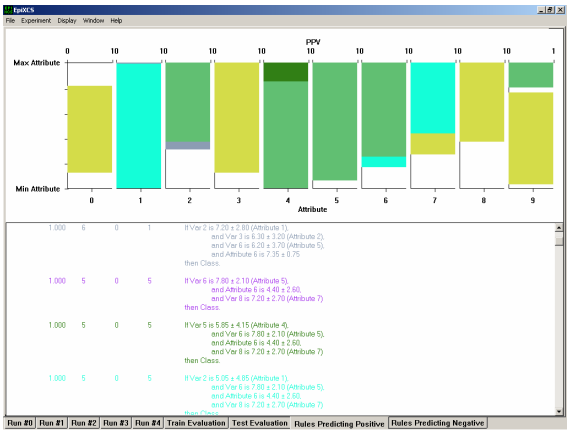

Figure 3. Ruleset derived by EpiXCS on the data with Odds Ratio=1.56 for Attribute 2.

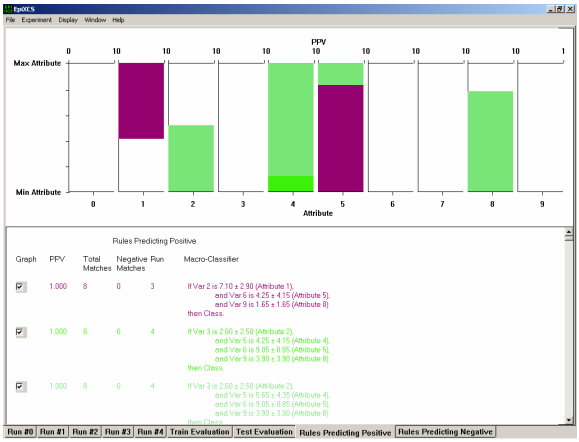

Figure 4. Ruleset derived by EpiXCS on the data with Odds Ratio of 2.90 for Attribute 2.

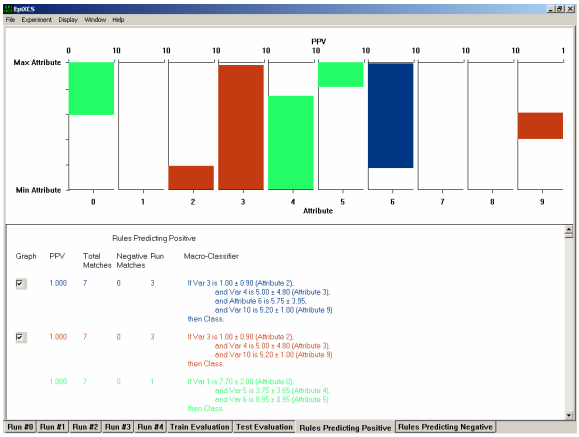

Figure 5. Ruleset derived by EpiXCS on the data with Odds Ratio of 3.41 for Attribute 2, for the positive class (cases).

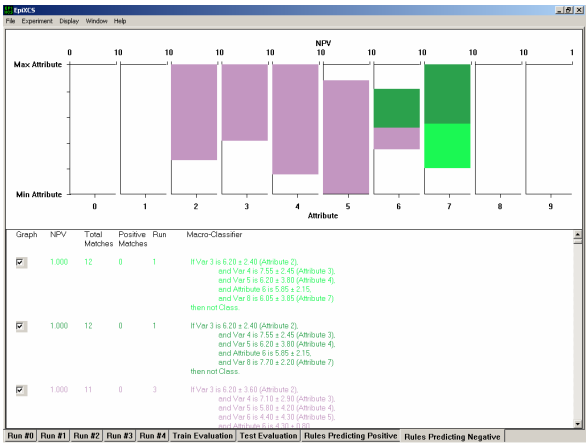

Figure 6. Ruleset derived by EpiXCS on the data with Odds Ratio=3.41 for Attribute 2, for the negative class (controls).

3.1.2 See5

See5 was unable to discover any rules on the baseline or the dataset with OR=1.56 for Attribute 2. Rules covering Attribute 2 for the positive class started to emerge at an OR of 2.06, but these rules were inaccurate, compared to the rules for the Controls. As was seen in EpiXCS, accurate rules which defined the range of values for Attribute 2, relative to class (0-2 for Cases and >2 for Controls), were not seen until the OR was to 3.41. In addition, these rules contained at least one other conjunct, as did those evolved by EpiXCS.

3.1.3 Logistic regression

While logistic regression does not discover "rules," per se, it does build a single model during supervised learning that reflects the relative contribution of each variable in the model to classifying the risk of an outcome. It is instructive to examine the ORs derived by logistic regression and their confidence intervals at each of the levels defined here, for two reasons. First, they provide the OR for the variable of interest, adjusted for every other variable in the model, simultaneously. This is important, given that all three methods used here are highly parallel in nature, and do consider individual variables simultaneously in model building. Second, the adjusted ORs provide a good sense of the complexity of the rule discovery problem. If logistic regression, as the "gold-standard" method finds statistically significant ORs at lower values for the variable of interest than do EpiXCS or See5, this would indicate an interesting inferiority of either, or both, of these methods.

The ORs and 95% confidence intervals for Attribute 2 are shown in Table 2. The ORs are approximately equivalent to the crude ORs obtained by simple 2x2 contingency table analysis. However, of more interest is the observation that the ORs are not statistically significant until the adjusted OR reaches 2.82, and even then only barely so. This indicates that logistic regression also had difficulty in discovering the association between low values of Attribute 2 and the positive class ("cases"). Finally, it should be noted that even where the ORs are significant, the confidence intervals are quite wide, which reflects the small size of the datasets, but also their complexity.

Table 2. Crude and adjusted Odds Ratios for the seven datasets.

Dataset	Crude (2x2 table)	Adjusted (logistic)
Baseline	1.00 (0.33, 3.00)	1.00 (0.32, 3.06)
1	1.56 (0.56, 4.42)	1.76 (0.62, 5.06)
2	2.06 (0.76, 5.73)	2.36 (0.85, 6.53)
3	2.45 (0.92, 6.75)	2.82 (1.04, 7.64)
4	2.90 (1.10, 7.92)	3.33 (1.23, 8.95)
5	3.41 (1.29, 9.28)	4.00 (1.51, 10.58)
6	4.00 (1.52, 10.86)	5.32 (1.93, 14.70)

4. DISCUSSION

This paper reports on an ongoing investigation into the use of EpiXCS as a knowledge discovery tool in epidemiologic surveillance. While the results presented here are preliminary, they indicate that EpiXCS is capable of discovering the values of specific attributes that are associated with a particular class in a supervised learning problem. The importance of this finding is substantial. Heretofore, logistic regression has been the analytic tool of choice for this problem. However, logistic regression analyses do not provide semantically useful rule ensembles that can be used easily for hypothesis generation. The single, mathematical model provided by logistic regression analysis is highly useful, but the advantage to the knowledge discovery process that is provided by the multiple disjunctions in a ruleset is lost.

The failure of See5 to discover any rulesets in the baseline or lower OR values for Attribute 2 is troubling, but expected. These datasets were quite equivocal; it was not until the OR for Attribute 2 had reached a substantial level, well over 2.0, that any meaningful rules emerged. Some comfort can be found in the observation that neither EpiXCS nor logistic regression performed well at this level. This could be a function of the small sample size, and this is a focus of ongoing study. It would be interesting to see if the suspected reason for poor performance by logistic regression (failure to converge due to small cell sizes) is similar in foundation to the reason for such failure by EpiXCS. Regardless, this remains an important issue in epidemiologic surveillance, and in order for these tools to be used fruitfully in this discipline, more research is needed to improve their ability to discover not only rare events, but also rare associations. It is the information contained in associational relationships that points clinical professionals to areas on which to intervene. Specifically, if the data used here were collected to monitor an outcome such as admission to the intensive care unit, and Attribute 2 measured a marker for exposure to an environmental toxin one would have to see a marked decrease in that marker before statistical significance (or rule emergence) was reached. It would be much better if the change in that marker could be identified much earlier as an association with an intensive care admission than it would have been in this study. To be more precise, an investigator would not have seen an important association between Attribute 2 and admission until nearly twice the number of cases with values of Attribute 2 <=2 had been observed. Depending on the actual situation, this could take a long time, and could cost considerably in terms of effort and potentially human life.

5. FUTURE DIRECTIONS

This is very much a nascent study, and much needs to be done in the area of applying learning classifier systems, particularly XCS (but others as well) to the problem of discovering attribute-class associations. First, larger, more complex datasets need to be developed and evaluated. The small size of the datasets used here represents a real limitation to interpreting the results of this study. Second, more work needs to be done with regard to parameterization. At smaller ORs, in the presence of more complex (read, conflicting) data, it is probably that population size, number of iterations, and perhaps genetic algorithm parameters will need to be adjusted. Finally, there is the intriguing possibility that the OR could be useful as a learning parameter. There could be at least two reasons why an OR would be low: the data don't support a higher OR, or the OR represents a rare, but emerging association. In the latter situation, the OR could be considered as an interestingness metric that would help drive reinforcement.

6. REFERENCES

[1] Butz, M.V. and Wilson S. W. An algorithmic description of XCS. In Lanzi, P. L., Stolzmann, W., and S. W. Wilson (Eds.), Advances in Learning Classifier Systems. Third International Workshop (IWLCS-2000), Lecture Notes in Artificial Intelligence (LNAI-1996). Berlin: Springer-Verlag (2001).

[2] Holmes, J.H.: The architecture of EpiXCS: An XCS-based learning classifier system for epidemiologic research. Sixth International Workshop on Learning Classifier Systems, Chicago, IL July 2003.

[3] Hulley S.B., et al: Designing Clinical Research. Philadelphia: Lippincott Williams and Wilkins, 2001, 122-123.

[4] Quinlan R: www.rulequest.com.

[5] STATA Corporation: STATA/SE 8.0 for Windows. College Station, TX.