From *Foundation and Recent Advances in Data Mining*, T.Y. Lin and C.J. Liau (eds.), 2005

A Careful Look at the Use of Statistical Methodology in Data Mining

Norman Matloff

Department of Computer Science
University of California, Davis
Davis, CA 95616 USA
matloff@cs.ucdavis.edu

Summary. Knowledge discovery in databases (KDD) is an inherently statistical activity, with a considerable literature drawing upon statistical science. However, the usage has typically been vague and informal at best, and at worst of a seriously misleading nature. In addition, much of the classical statistical methodology was designed for goals which can be very different from those of KDD. The present paper seeks to take a first step in remedying this problem by pairing precise mathematical descriptions of some of the concepts in KDD with practical interpretations and implications for specific KDD issues.

1 Introduction

The field of KDD has made extensive use of statistical methodology. Such methodology is clearly of great potential, but is also fraught with a myriad of pitfalls. A lack of insight into how the methods actually work may result in unnecessarily weak KDD machinery. Moreover, naive "transplantation" of many statistical methods to KDD arenas for which the methods were not designed may result in poor or even misleading analyses.

The remedy is for KDD practitioners to on the one hand to gain a better, more precise mathematical understanding of the statistical methodology, and on the other hand to develop a better intuitive understanding of what the methodology does.

In this paper, we will encourage KDD practitioners to

- devise simple mathematical models which will facilitate precise statements of the problems at hand and deepen intuitive insight into possible solutions

- take a close look at the goals of the statistical methodology they use, assessing how well those methods fit the given KDD application

We will present a simple framework, consisting of some simple mathematical constructs motivated by intuitive notions tied to the actual practice of KDD. It is important to note that the latter, i.e. the intuitive "philosophical" issues, will play an integral role here.

It is assumed here that the reader has at least a first-level knowledge of standard statistical methods, e.g. hypothesis testing, confidence intervals and regression, and a basic background in probabilistic constructs such as random variables, expected value and so on. We first develop some simple infrastructure, some of which will be at least somewhat familiar to many readers, and then move to detailed worked-out examples which illustrate the issues.

2 Statistical Sampling

As is common in theoretical treatments, we will phrase the issues in terms of a statistical prediction problem. This is not to say we consider KDD to be limited to prediction settings, but rather that such settings are among the most common KDD applications. We depart from tradition, though, by engaging in an explicit discussion of the practical interpretation of what we mean by "statistical."

2.1 Notation

Denote our attribute set by $X^{(1)}, ..., X^{(d)}$. It is assumed that our database constitutes a *statistical sample* of n observations on these attributes; the i^{th} observation on the j^{th} attribute from this sample is denoted by $X_i^{(j)}$, i = 1,...,n, j = 1,...,d. We need to spend some time here on the question of what this really means.

To make things concrete—again, this is one of our principle aims—let's consider the well-known KDD "market basket" example. Each row in the database corresponds to some individual consumer. The j^{th} attribute might record whether the consumer bought a given item (1 for yes, 0 for no).[1] We wish to predict the event $X^{(i)} = 1$ from several other simultaneous events $X^{(j)} = 1$. In other words, we wish to know whether a consumer's purchase of one item is related to his/her possible interest in buying a related item. For example, in an online book sales site, if a consumer purchases a certain book, the site may then suggest to the customer that he/she consider buying other related books. Note that this is unlike typical statistical contexts, in which we would be interested in predicting from events of the form $X^{(j)} = 0$ as well.

[1] In other KDD contexts, some of the attributes may be characteristics of a consumer, say age, income or gender (say 1 for male, 0 for female), and so on.

The vector $(X_i^{(1)}, ..., X_i^{(d)})$, representing the values of all our attributes in the i^{th} observation will be denoted by X_i. In relational database terms, this vector is the i^{th} row in our relation.

2.2 Sampling from Populations, Real or Conceptual

In considering our database to be a "statistical sample," we mean that it is a sample from some "population." This interpretation is, in our view, key.

The population may be tangible, as in the "basket" example, where we are sampling from the population of all customers of this business. Or, the population may be more conceptual in nature. A database consisting of students in a new curriculum in a university could be considered as a sample from the conceptual population of all students at this university who *might* be in this major. If for example we imagine the university overall enrollment had been 20 percent larger this year, with no change in demographic or other makeup of the enrollment, then some of the increased overall enrollment would have been students choosing this major. This population is then conceptual. Similar remarks hold when considering potential future students in the major.

Here is an example of a "population" which is even more conceptual in nature. Consider the subject of quadratic equations, studied in school algebra classes:

$$ax^2 + bx + c = 0 \tag{1}$$

The students learn that this equation has a real root if and only the *discriminant* $b^2 - 4ac$ is nonnegative. Suppose one did not know this rule, and tried to find it using KDD.

This sounds like an inherently non-statistical problem. Yet one could convert it to a statistical problem in the following way. One could sample randomly from $a/b/c$ space, according to a distribution of one's choice, and for each sample triplet from this space, determine somehow (say by graphing the quadratic polynomial) whether a real root exists. One could then apply various *statistical regression models* (see below), trying to predict the 0-1 variable w from a, b and c, where w is 1 if there are real roots and 0 otherwise. In this manner, we might possibly stumble onto the discriminant rule.

2.3 Relation to Probability Distributions

It is important to relate the abstract mathematical variables to the population being studied. When we speak of the distribution of the random variable $X^{(j)}$, what we really mean is the distribution of that attribute in the population. Say $X^{(1)}$ is age of the customer. When we say, for instance, that $P(X^{(1)} > 32) = 0.22$, we mean that 22 percent of all customers in this population are older than 32.

A similar point holds for expected value. Some KDD practitioners with an engineering or physical science background might be accustomed to interpreting $E(X^{(1)})$ in terms of the physics metaphor of center of gravity. Yet for statistical applications such as KDD, the relevant interpretation of this quantity is as the mean age of all customers in this population. This interpretation is especially important when considering sample-distribution issues such as bias and variance, as we will see.

3 Prediction

As we have noted, our focus in the statistical nature of KDD is on prediction. For notational convenience, in the remainder of this paper, let us suppose that we are using $X^{(1)}, ..., X^{(d-1)}$ to predict $X^{(d)}$, and rename the latter attribute Y.

Our focus here will largely be on predicting *dichotomous*, i.e. 0/1-valued, attributes Y. (We do not make this restriction on the attributes $X^{(j)}$.) However, as will be seen, most of the issues which arise also occur in the case of continuous-valued Y.

3.1 Statement of the Problem

Suppose for the moment that we know the population distributions of the attributes, and we wish to minimize the overall probability of misclassification.[2] Suppose that we observe $X^{(j)}$ to have the value v_j, j = 1,...,d-1. Then we would guess Y to be either 0 or 1, according to whether the quantity

$$\begin{aligned} r(v) &= r(v_1, ..., v_{d-1}) \\ &= P(Y = 1 | X^{(1)} = v_1, ..., X^{(d-1)} = v_{d-1}) \end{aligned} \tag{2}$$

is less than 0.5 or greater than 0.5, respectively, where $v = (v_1, ..., v_{d-1})$).[3] Let us then denote this guess as g[r(v)], where g(u) = floor(2u) for u in [0,1].

Note that in theory r(v) should converge to 1 or 0 (depending on v) as $d \rightarrow \infty$. In other words, if you know enough about the situation, you can always predict correctly! This of course is a very big “if”, but it puts in perspective notions such as that of *unexpected rules* in [10]. Nothing is “unexpected,” strictly speaking; we simply lack data. This issue will become relevant in our discussion of the bias/variance tradeoff and Simpson's Paradox later.

[2] The latter would not be the case if we assigned different costs to different types of errors. It may be more costly to falsely guess Y to be 1 than to falsely guess it to be 0, for example.

[3] Here and below, will write this function simply as r(), suppressing the dependence on d, i.e. not writing it as r_{d-1}. The dependence will be clear from the number of arguments.

3.2 Classification Vs. Regression

Some authors, e.g. Han [5] consider the case of dichotomous Y to be a conceptually separate case from that of continuous Y, and refer to it as *classification* instead of prediction. However, mathematically it is the same problem, in the following sense.

Classically, the problem of predicting a general attribute Y from a vector of attributes $X = (X^{(1)}, ..., X^{(d-1)})$ is posed as finding a function h() that minimizes

$$E[(Y - h(X))^2] \tag{3}$$

One can easily show that the minimizing solution is the *regression function*,

$$h(t) = E(Y|X = t) \tag{4}$$

Now, if Y is dichotomous, i.e. Y takes on the values 0 and 1, then

$$E(Y|X = t) = w \cdot 1 + (1 - w) \cdot 0 = r(t) \tag{5}$$

where w = P(Y=1|X=t).

In other words, the general formulation of the prediction problem yields the regression function r() anyway.

Thus the classification and regression problems are the same. This is not just a semantic issue. There are indeed some aspects of the classification problem which differ from the regression setting, but there is a great amount of commonality. A vast literature exists on the general regression problem, with much material relevant to the dichotomous case,[4] and it would be a loss not to draw upon it.[5]

3.3 The Function r() Must Be Estimated from Sample Data

The situation is complicated by the fact that we do <u>not</u> know the population distributions of the attributes, as assumed in the previous few paragraphs. We thus do not know the function $r()$ above, and need to estimate it from the observations in our database.

The estimated function, $\hat{r}(v)$, is obtained either by parametric or nonparametric means. A common parametric approach, for instance, uses the logistic regression model, which postulates that $r(v)$ has the form

$$r(v_1, ..., v_{d-1}) = \frac{1}{1 + exp[-(\beta_0 + \beta_1 v_1 + ... + \beta_{d-1} v_{d-1})]}$$

[4] A large separate literature on the classification problem has also been developed, but much of it draws upon the material on regression.

[5] By the way, some of these points are also noted (albeit rather abstractly) in [3].

The parameters β_j are estimated from our sample data $X_i^{(j)}$, yielding the estimated parameters $\hat{\beta}_j$ and the estimated r(v):[6]

$$\hat{r}(v) = \frac{1}{1 + exp[-(\hat{\beta}_0 + \hat{\beta}_1 v_1 + ... + \hat{\beta}_{d-1} v_{d-1})]}$$

Many nonparametric method have been used in KDD for estimating r(v), such as CART [1].

4 Over/underfitting

There is a lot of talk about "noise mining," "overfitting" and the like in the KDD literature, but again this is rarely precisely defined.

4.1 Bias and Variance

In some theoretical papers, the literature does at least point out that the "average" discrepancy between $\hat{r}(v)$ and $r(v)$ can be shown to consist of two components—(the square of) a *bias* component,

$$E\hat{r}(v) - r(v) \tag{6}$$

and a *variance* component,

$$E[(\hat{r}(v) - E\hat{r}(v))^2] \tag{7}$$

Note that v is fixed here, not random. Instead, the randomness involves the fact that these expected values are averages over all possible n-observation samples from the given population.[7] This interpretation is very important when one is assessing various competing types of prediction methodology, and especially important in understanding the bias/variance problem.

A large bias is due to using too simple a model in the parametric case, or to using too coarse a granularity in the nonparametric case (e.g. leaf nodes too large in CART). In both cases, one common source of the problem is that we are using too few predictor attributes.

However, any efforts to reduce the bias will increase the variance, i.e. increase the amount of "noise." This is due to having an insufficient sample size n for the given model. In CART, for example, if we use smaller hyper-rectangles in order to reduce bias, a given hyper-rectangle might contain very few observations, thus causing $\hat{r}()$ to have a large variance within the rectangle. The same rectangle, applied to a larger sample from the same population,

[6] See for example the **lrm** procedure in the R statistical package [11].

[7] We could also make v random. i.e. replace it by X in the expressions above, so that the averaging is being done both over all possible n-observation samples and over all values of v.

might work fine. In the case of a logistic model, if we add more predictor variables in order to reduce bias, then for each fixed j, $Var(\hat{\beta}_j)$ will tend to increase.[8]

These phenonema arise in the "market basket" setting as well. If confidence and support levels (see Section 6.1) are set too low, a bias problem occurs in much the same way as it does if we have too large a rectangle in CART. If the confidence and support levels are set too high, it becomes a variance problem.

4.2 Illustrative Model

Recall that our theme here has been that empirical research work in KDD should include a mathematically precise statement of the problem, and should present mathematical treatment of at least a small but illustrative model of the effects being studied. In that light, we now present such a model of the "noise fitting" problem. We ask the reader to keep in mind our specific goal here in devising this model—we desire to devise a simple but precise model in which the roles of both d and n in the bias/variance tradeoff are explicitly visible in the model's results.[9] It is our hope that KDD practitioners will often engage in setting up such simple models in order to gain insight into specific applications.

Continue to assume the setting described at the beginning of Section 3, but with the additional specialization that all the predictor attributes $X^{(j)}$, j = 1,...,d-1 are dichotomous.

Suppose that $X^{(j)}$, j = 1,...,d-1 are all "coin tosses," i.e. have probability 0.5 of taking on the value 1 and are statistically independent. Suppose in addition that $P(Y = 1|X^{(1)} = v_1)$ is equal to 0.6 for $v_1 = 1$ and equal to 0.4 for $v_1 = 0$, and that $X^{(j)}$, j = 2,...,d-1 have no predictive power for Y at all, i.e.

$$r(v_1, v_2, ..., v_{d-1}) = P(Y = 1|X^{(1)} = v_1) \tag{8}$$

independent of $v_2, ..., v_{d-1}$.

But we would not have this information, since we would not have the population data. We would have only *sample* estimates of r(v) to work with, $\hat{r}(v)$. The point then is that that estimate will be subject to bias/variance issues. We discuss the variance issue first, and focus our attention on the estimation of r(1,1,...,1).

One decision we would need to make is which of the attributes $X^{(j)}$ to use as predictors. Let us compare the effects of using just $X^{(1)}$ alone to predict Y, versus using $X^{(1)}, X^{(2)}, ..., X^{(d-1)}$ for that prediction. In the former situation, note again that we would be modeling r(v) to be a function which

[8] See [14] for an analytical proof of this in the linear regression setting.

[9] A number of much more complex examples of this tradeoff for various kinds of estimators are presented (without derivation) in [6].

does not depend on $X^{(2)}, ..., X^{(d-1)}$ (see Equation (8)). Again, this modeling assumption would be correct, but we would not know this.

Suppose we are not using a parametric model, and instead are simply using straight sample proportions to estimate r(). Then if we use only $X^{(1)}$ as our predictor, our estimate of r(1,1,...,1) would be the proportion of records in our database for which Y = 1, among those for which $X^{(1)} = 1$, i.e.

$$\hat{r}(1,1,...,1) = \frac{\sum_i X_i^{(1)} X_i^{(d)}}{\sum_i X_i^{(1)}} = \frac{T_1}{U_1} \tag{9}$$

Recalling that r(1,1,...,1) = 0.6, the question at hand is, "What is the probability that $\hat{r}(1,1,...,1)$ will make the right decision for us in this situation, which is to guess that Y = 1?"[10] Well, this is

$$P(\hat{r}(1,1,...,1) > 0.5) = P(T_1 > 0.5U_1) \tag{10}$$

To evaluate this probability, note first that T_1 and U_1, are binomially distributed.[11] Thus they have approximate normal distributions. But in addition, their bivariate distribution approximates that of a bivariate normal.[12] The means and variances of T_1 and U_1 are then np, nq, np(1-p) and nq(1-q), where $p = P(X^{(1)} = X^{(d)} = 1) = 0.3$ and $q = P(X^{(1)} = 1) = 0.5$. Their covariance is

$$\begin{aligned} & Cov(T_1, U_1) \\ & = n[E(X^{(1)} X^{(d)} X^{(1)}) - E(X^{(1)} X^{(d)}) \cdot EX^{(1)}] \\ & = np(1-q) \end{aligned} \tag{11}$$

Any linear combination of T_1 and U_1, say $aT_1 + bU_1$, then has an approximate normal distribution with mean n(ap+bq), and variance

$$n[a^2 Var(T_1) + b^2 Var(U_1) + 2abCov(T_1, U_1)] \tag{12}$$

In our case here, a = 1 and b = -0.5. After doing the calculations we find that $E(T_1 - 0.5U_1) = 0.05n$ and $Var(T_1 - 0.5U_1) = 0.1225n$, and thus

$$P(T_1 > 0.5U_1) \approx 1 - \Phi(-0.14\sqrt{n}) \tag{13}$$

where Φ is cumulative distribution function of a standard N(0,1) variate.

[10] Note that the term "the right decision" means the decision we would make if we had full knowledge of the population distributions, rather than just sample estimates. It does not mean that our guess for Y is guaranteed to be correct.

[11] The variable $B = X^{(1)} X^{(d)}$ is 0-1 valued, and the terms are independent, so the sum T_1 is binomial.

[12] This stems from the fact the vector form of the Central Limit Theorem.

So, Equation (13) is the probability that we make the right decision if we predict Y from only $X^{(1)}$. Let's see how that probability changes if we predict Y from $X^{(1)}, ..., X^{(d-1)}$.

In this setting, Equation (2) again reverts to (8), and (9) becomes

$$\hat{r}(1, 1, ..., 1) = \frac{\sum_i X_i^{(1)} X_i^{(2)} ... X_i^{(d-1)} X_i^{(d)}}{\sum_i X_i^{(1)} X_i^{(2)} ... X_i^{(d-1)}} = \frac{T_{d-1}}{U_{d-1}} \tag{14}$$

The analog of (13) is then (after a bit of algebraic approximation)

$$P(T_{d-1} > 0.5U_{d-1}) \approx 1 - \Phi\left(-0.28 \cdot \sqrt{\frac{n}{2^d}}\right) \tag{15}$$

Compare Equations (13) and (15), focusing on the roles of d and n. They are both of the form $P(Z > c)$ for a negative c, and the algebraically smaller (i.e. more negative) c is, the better. So, for fixed n, the larger d is, the worse is our predictive ability for Y, if we use all d-1 predictors.

Now, remember the context: We devised a model here in which $X_i^{(2)} ... X_i^{(d-1)}$ had no predictive ability at all for Y in the population distribution, though the analyst would not know this. *In other words, not only will the analyst not gain predictive ability by using these attributes, he/she would actually lose predictive power by using them, i.e. "overfit."*

So, this is the variance side of the bias/variance tradeoff. The number of records in our sample which have $X^{(1)} = 1, X^{(2)} = 1, ..., X^{(d-1)} = 1$ will be very small for large d (similar to having a small leaf node in CART), leading to a high variance for $\hat{r}(1, 1, ..., 1)$.[13]

Equation (15) also shows the role of n in the overfitting issue: For fixed d, as n increases the harmful effect of overfitting will diminish.

Now, what about the bias side of the bias/variance tradeoff? Suppose we are considering using $X^{(1)}, X^{(2)} ..., X^{(k)}$ as our predictor attributes. Due to the nature of the model here, i.e. the fact that $X^{(2)} ..., X^{(k)}$ have no predictive power, the bias in using any k in the range $1 \leq k < d - 1$ is 0.[14] So, if we use k greater than 1, we are incurring the problem of increasing variance without reducing bias, a pure loss.

On the other hand, using k = 0 would produce a bias, since $X^{(1)}$ does have some predictive value for Y: If k were taken to be 0, then the population value of r(1,1,...,1) would reduce to the unconditional probability P(Y = 1) = 0.5, rather than the achievable value 0.6.

[13] We did not directly calculate that variance here, showing the variance effects only indirectly. However, $Var(\hat{r})$ could be calculated by using the *delta method*, applied to the function f(t,u) = t/u [13].

[14] Here we are using the term *bias* only informally, not quite as in Equation (6). To make it formal, we would have to make the expected value conditional on U_k, the latter defined analogously to U_1 and U_{d-1} for the case of k predictors. The technical reason for this is that U_k may be 0, and thus $\hat{r}(1, 1, ..., 1)$ would be undefined.

Again, our point in devising this model here is to illustrate our theme that even empirical KDD research should anchor its presentation with (a) a precise mathematical statement of the problem being studied, and (b) a simple mathematical model which explicitly illustrates the issues.

The word *explicitly* in (b) should be emphasized. Equation (15) explicitly shows the roles of d and n. One sees that for a fixed value of n, use of a larger d increases the variance. As d increases, at some point our predictive ability based on sample data will begin to diminish, i.e. we will overfit. One also sees, though, that for a larger value of n, that crossover point will occur for a larger d, i.e. we can use more attributes as our predictors.

5 Attribute Selection

As we have seen, there is a tradeoff between bias and variance for fixed n. As finer models are fitted, with more attributes, the bias is reduced (or stays the same) but the variance increases. If too much attention is paid to minimizing bias rather than variance, the decision rules found from the analysis may be spurious, hence the term *noise mining.*

The problem of finding the attribute set which maximizes predictive ability, i.e. finding the optimal point in the bias/variance tradeoff spectrum, is as old as the field of statistics itself. It must be emphasized in the strongest possible terms that this is still an unsolved problem, in spite of numerous papers in the KDD literature which report on "promising" solutions.

5.1 The Use of Hypothesis Testing

We wish to emphasize also the importance of phrasing the problem directly in terms of the goals of the KDD settings being analyzed. For example, the classical statistical method for selecting predictor attributes, hypothesis testing, is of questionable propriety. In the case of a logistic model, say, this approach would involve testing the hypothesis

$$H_0 : \beta_j = 0 \tag{16}$$

and then either including or excluding the attribute $X^{(j)}$ in our predictor set, depending on whether the hypothesis is rejected or accepted. This procedure is often applied sequentially, one potential predictor attribute at a time, to determine which attributes to use; this algorithm is called *stepwise variable selection.*

Yet the classic use of hypothesis testing in regression analysis is largely aimed at descriptive, rather than predictive, types of applications. An example of descriptive use is the identification of risk factors for a certain disease. The attributes having large β_j are considered to be the more important factors. There the goal is more to understand how a disease arises in the population at large, rather than to actually predict whether a particular individual develops the disease. By contrast, in many KDD settings one really does want to predict, and thus one should be hesitant to apply classical variable-selection algorithms in the KDD context.

For example, in the classical use of regression, the hypotheses are typically tested at traditional significance levels such as $\alpha = 0.05$. Yet some studies (e.g. [7]) have found that for the prediction of continuous variables, the best values of α are in the range 0.25 to 0.40.[15]

Some readers will notice this as a statistical *power* issue; the term refers to the probabilities of rejecting the null hypothesis under various scenarios of the alternative hypothesis. As such, it is not a new concept, since the original usage of hypothesis testing in the early 20th century did assume that analysts would carefully balance the values of α and power in a manner suitable to the application. However, modern usage has institutionalized the value of α to be set to 0.05 or 0.01, so much so that a "star" notation has become standard in research in the social and life sciences (a statistic is adorned with one or two asterisks, depending on whether α is 0.05 or 0.01).[16] Power is rarely discussed, let alone calculated.

Again, in KDD the goals may be very different from classical statistical usage, and thus that our analyses must not blindly mimic that usage. In this case, the point is that if one does use hypothesis testing for model selection, power considerations are crucial. Of course, it is not always clear how best to use power analyses in a given situation, or even how to calculate it in many cases, but it is certainly clear that classical values of α are not the best.

6 The Multiple Inference Problem

If hypothesis testing is used for attribute selection, there is not only the problem of considering power levels, but also the issue of accuracy of the α and power levels. First there is the problem that even though each test has the specified significance level, the collective significance level of all the tests may be much greater. In addition, many attribute selection algorithms, e.g. stepwise selection, are *adaptive*, and thus even the individual significance levels may have values quite different from their nominal values.

The problem of the collective significance level being much greater than the level applied to each individual test can be addressed by the use of *multiple inference methods*, which allow one to set an overall significance level for multiple tests. In this light, a worthy future research project would be to revisit past dismissals of the use of multiple inference for rule finding [8].

[15] The picture is further muddled by the fact that the stated α value is nominal anyway, due to issues such as *multiple inference*, discussed in Section 6.

[16] Even if one's goal is descriptive rather than predictive, the usage of these institutionalized values is questionable [9].

Earlier authors had found such approaches to be too conservative, finding too few rules. However, the point we made above in Section 5.1 suggests that with a larger value of overall α, such methodology may work well.

6.1 Illustrative Example

Consider the market basket problem with two attributes, $X^{(1)}$ and $X^{(2)}$. Typically one is interested in finding all attributes for which the *confidence*, say

$$P(X^{(2)} = 1|X^{(1)} = 1) \tag{17}$$

is greater than a given level c, and the *support*, say

$$P(X^{(2)} = X^{(1)} = 1) \tag{18}$$

is a above s. If both conditions are satisfied, we will use the rule $X^{(1)} \Rightarrow X^{(2)}$.

(For the sake of simplicity, we are using only two attributes in this example. Typically there are many more than two attributes, in which case combinations of attributes are considered. With three attributes, for instance, we would assess not only potential rules such as $X^{(1)} \Rightarrow X^{(2)}$ but also some like $X^{(1)}, X^{(3)} \Rightarrow X^{(2)}$. In the latter case, quantities such as $P(X^{(2)} = 1|X^{(1)} = X^{(3)} = 1)$ would be checked.)

Let $p_{ij} = P(X^{(1)} = i, X^{(2)} = j)$, i,j = 0,1. Then to determine whether to use the rule $X^{(1)} \Rightarrow X^{(2)}$, we might test the hypothesis

$$H_0 : p_{11} \leq s \ or \ \frac{p_{11}}{p_{11} + p_{10}} \leq c \tag{19}$$

and then use the rule if the hypothesis is rejected. But for mathematical tractability here, let us treat (19) as two separate hypotheses. Accounting also for the possible rule $X^{(2)} \Rightarrow X^{(1)}$, we have a total of three hypotheses to test in all.[17] We will now investigate how well we can assess the two rules with a given value of α. We will calculate E(K), where K is the number of hypotheses in which we make the correct decision; K varies from 0 to 3.

As a test case, let us take the matrix $p = (p_{ij})$ to be

$$\begin{pmatrix} 0.15 & 0.45 \\ 0.10 & 0.30 \end{pmatrix} \tag{20}$$

and take s = 0.35, c = 0.60. In this setting, the potential rules have confidence and support as shown in Table 1. Then in this simple example,[18] of the three hypothesis tests to be performed, ideally two of them should be rejected and one accepted. E(K) will be the sum of the probabilities of the two rejections and one acceptance.

A standard method for multiple inference on a small number of tests involves use of the *Bonferroni Inequality.*[19] If one is performing k tests and

[17] The two potential rules have one support test in comon.

[18] A much more general study is being conducted for a separate paper.

[19] A reference on this and many other multiple inference methods is [12].

poss. rule	conf.	supp.
$X^{(1)} \Rightarrow X^{(2)}$	0.75	0.30
$X^{(2)} \Rightarrow X^{(1)}$	0.40	0.30

Table 1. Rule Results

wishes an overall significance level of at most α, then one sets the individual significance level of each test at α/k. Here, to achieve an overall significance level of 0.05, we use a level of $0.05/3 = 0.017$ for each of the three tests. For a one-sided test, this corresponds to a "Z value" of 2.12 in the normal distribution, i.e. $1 - \Phi(2.12) = 0.017$.

To test the potential rule $X^{(1)} \Rightarrow X^{(2)}$ for our confidence level 0.35, we reject if

$$\frac{\hat{p}_{11} - 0.35}{\sqrt{\hat{Var}(\hat{p}_{11})}} > 2.12 \tag{21}$$

where

$$\hat{Var}(\hat{p}_{11}) = \frac{\hat{p}_{11}(1 - \hat{p}_{11})}{n} \tag{22}$$

Thus we need to compute

$$P\left(\frac{\hat{p}_{11} - 0.35}{\sqrt{\hat{Var}(\hat{p}_{11})}} < 2.12\right) \tag{23}$$

in the setting $p_{11} = 0.30$. This probability is computed (approximately) via standard normal distribution calculations as seen earlier in Section 4.2.[20]

For testing whether the potential rule $X^{(1)} \Rightarrow X^{(2)}$ meets the confidence criterion, note that

$$\frac{p_{11}}{p_{11} + p_{10}} > c \tag{24}$$

holds if and only if

$$(1 - c)p_{11} - cp_{10} > 0 \tag{25}$$

Thus the probability of making the correct decision regarding the confidence level of $X^{(1)} \Rightarrow X^{(2)}$ is

[20] Here, though, one uses the exact variance in the denominator in (23), i.e. $p_{11}(1 - p_{11})/n = 0.18/n$.

$$P\left(\frac{(1-c)\hat{p}_{11}-c\hat{p}_{10}}{\sqrt{\hat{Var}[(1-c)\hat{p}_{11}-c\hat{p}_{10}]}}>2.12\right) \tag{26}$$

where, using (12) and the fact that $Cov(\hat{p}_{11},\hat{p}_{10})=-p_{11}p_{10}/n$,

$$\sigma^2=Var[(1-c)\hat{p}_{11}-c\hat{p}_{10}]=$$

$$\frac{1}{n}[(1-c)^2p_{11}(1-p_{11})+c^2p_{10}(1-p_{10})+2c(1-c)p_{11}p_{10}]$$

The probability of the correct decision is then

$$1-\Phi\left(2.12-\frac{(1-c)p_{11}-cp_{10}}{\sigma}\right) \tag{27}$$

The probability of the correct decision for $X^{(2)}\Rightarrow X^{(1)}$ (that the confidence level is *not* met) is computed similarly.

After doing all this computation, we found that the value of E(K) for an α of 0.05 turns out to be 2.70. After some experimentation, we found that a very large level of 0.59 improves the value of E(K) to 2.94. Though this amount of improvement is modest, it does show again that the best choice of significance level in KDD settings may be quite different from those used classically in statistical applications. Moreover, it shows that multiple inference methods may have high potential in KDD after all, if only one considers nontraditional significance levels. And as mentioned, the multiple-inference approach dismissed in [8] appear to be worth revisiting.

The fact that the relevant hypotheses involve linear combinations of the p_{ij}, as in (25), suggests that the Scheffe' method of multiple inference could be used. That method can simultaneously test all linear combinations of the p_{ij}. Those rules for which the confidence and support tests are both rejected would be selected. Again, it may be the case that with suitable significance levels, this approach would work well. As noted earlier, this is under investigation.

7 Simpson's Paradox Revisited

A number of KDD authors have cautioned practitioners to be vigilant for *Simpson's Paradox* [4]. Let us first couch the paradox in precise mathematical terms, and then raise the question as to whether, for predictive KDD settings, the "paradox" is such a bad thing after all.

Suppose each individual under study, e.g. each customer in the market basket setting, either possesses or does not possess traits A, B and C, and that we wish to predict trait A. Let $\bar{A}$, $\bar{B}$ and $\bar{C}$ denote the situations in which the individual does not possess the given trait. Simpson's Paradox then describes a situation in which

$$P(A|B) > P(A|\bar{B}) \tag{28}$$

and yet

$$P(A|B,C) < P(A|\bar{B},C) \tag{29}$$

In other words, the possession of trait B seems to have a positive predictive power for A by itself, but when in addition trait C is held constant, the relation between B and A turns negative.

An example is given in [2], concerning a classic study of tuberculosis mortality in 1910. Here the attribute A is mortality, B is city (Richmond, with $\bar{B}$ being New York), and C is race (African-American, with $\bar{C}$ being Caucasian). In probability terms, the data show that:[21]

- P(mortality | Richmond) = 0.0022
- P(mortality | New York) = 0.0019
- P(mortality | Richmond, black) = 0.0033
- P(mortality | New York, black) = 0.0056
- P(mortality | Richmond, white) = 0.0016
- P(mortality | New York, white) = 0.0018

The data also show that

- P(black | Richmond) = 0.37
- P(black | New York) = 0.002

a point which will become relevant below.

At first, New York looks like it did a better job than Richmond. However, once one accounts for race, we find that New York is actually worse than Richmond. Why the reversal? The answer stems from the fact that racial inequities being what they were at the time, blacks with the disease fared much worse than whites. Richmond's population was 37% black, proportionally far more than New York's 0.2%. So, Richmond's heavy concentration of blacks made its overall mortality rate look worse than New York's, even though things were actually much worse in New York.

But is this “paradox” a problem? Some statistical authors say it merely means that one should not combine very different data sets, in this case white and black. But is the “paradox” really a problem in KDD contexts?

The authors in [2] even think the paradox is something to be exploited, rather than a problem. Noting that many KDD practitioners are interested in finding “surprising” rules (recall [10]), the authors in [2] regard instances of Simpson's Paradox as generally being surprising. In other words, they contend that one good way to find surprising rules is to determine all instances of Simpson's Paradox in a given data set. They then develop an algorithm to do this.

[21] These of course are sample estimates.

That is interesting, but a different point we would make (which, to our knowledge has not been made before) is that the only reason this example (and others like it) is surprising is that the predictors were used in the wrong order. As noted in Section 5, one normally looks for predictors (or explanatory variables, if the goal is understanding rather than prediction) one at a time, first finding the best single predictor, then the best pair of predictors, and so on. If this were done on the above data set, the first predictor variable chosen would be race, not city. In other words, the sequence of analysis would look something like this:

- P(mortality | Richmond) = 0.0022
- P(mortality | New York) = 0.0019
- P(mortality | black) = 0.0048
- P(mortality | white) = 0.0018
- P(mortality | black, Richmond) = 0.0033
- P(mortality | black, New York) = 0.0056
- P(mortality | white, Richmond) = 0.0016
- P(mortality | white, New York) = 0.0018

The analyst would have seen that race is a better predictor than city, and thus would have chosen race as the best single predictor. The analyst would then investigate the race/city predictor pair, and would never reach a point in which city alone were in the selected predictor set. Thus no anomalies would arise.

8 Discussion

We have, in the confines of this short note, endeavored to argue for the need for more mathematical content in empirically-oriented KDD research. The mathematics should be kept simple and should be carefully formulated according to the goals of the research. We presented worked-out examples of how a simple mathematical model could be used to illustrate, and hopefully gain insight into, the issues at hand.

We wish to reiterate, on the other hand, that very theoretical treatments, written by and for mathematical statisticians, are generally inaccessible to empirical KDD researchers and KDD practitioners. We hope that theoretical work be made more intuitive and tied to practical interpretations.

References

1. L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth Publishers, 1984.

2. C.C. Fabris and A.A. Freitas. Discovering Surprising Patterns by Detecting Occurrences of Simpson's Paradox. In *Research and Development in Intelligent Systems XVI (Proc. ES99, The 19th SGES Int. Conf. on Knowledge-Based Systems and Applied Artificial Intelligence)*, 148-160. Springer-Verlag, 1999.
3. J. Friedman. On Bias, Variance, 0/1-Loss, and the Curse of Dimensionality, *Data Mining and Knowledge Discovery*, 1, 55-77, 1997.
4. C. Glymour, D. Madigan, D. Pregibon and P. Smyth. Statistical Themes and Lessons for Data Mining, *Data Mining and Knowledge Discovery*, 1, 25-42, 1996.
5. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*, Morgan Kaufman Publishers, 2000.
6. T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
7. R. Bendel and A. Afifi, Comparison of Stopping Rules in Forward Stepwise Regression, Joint ASA/IMS Meeting, St. Louis, 1974.
8. P. Domingos. E4—Machine Learning. http://citeseer.nj.nec.com/205450.html.
9. N. Matloff, Statistical Hypothesis Testing: Problems and Alternatives. *Journal of Economic Entomology*, 20, 1991, 1246-1250.
10. B. Padmanabhan and A. Tuzhilin, Finding Unexpected Patterns in Data, in *Data Mining, Rough Sets and Granular Computing*, T.Y. Lin, Y.Y. Yao, L. Zadeh (eds), Physica-Verlag, 2001.
11. R statistical package home page, http://www.r-project.org.
12. Y. Hochberg and A. Tamhane, *Multiple Comparison Procedures*, Wiley, 1987.
13. R.J. Serfling, *Approximation Theorems of Mathematical Statistics*, Wiley, 1980.
14. R. Walls and D. Weeks. A Note on the Variance of a Predicted Response in Regression, *The American Statistician*, 23, 24-26, 1969.