Multiple Hypothesis Testing for Data Mining

Sach Mukherjee
University of Oxford, UK

INTRODUCTION

A number of important problems in data mining can be usefully addressed within the framework of statistical hypothesis testing. However, while the conventional treatment of statistical significance deals with error probabilities at the level of a single variable, practical data mining tasks tend to involve thousands, if not millions, of variables. This Chapter looks at some of the issues that arise in the application of hypothesis tests to multi-variable data mining problems, and describes two computationally efficient procedures by which these issues can be addressed.

BACKGROUND

Many problems in commercial and scientific data mining involve selecting objects of interest from large datasets on the basis of numerical relevance scores ("object selection"). This Section looks briefly at the role played by hypothesis tests in problems of this kind. We start by examining the relationship between relevance scores, statistical errors and the testing of hypotheses in the context of two illustrative data mining tasks. Readers familiar with conventional hypothesis testing may wish to progress directly to the main part of the Chapter.

As a topical example, consider the differential analysis of gene microarray data (Piatetsky-Shapiro & Tamayo, 2004; Cui & Churchill, 2003). The data consist of expression levels (roughly speaking, levels of activity) for each of thousands of genes across two or more conditions (such as healthy and diseased). The data mining task is to find a set of genes which are differentially expressed between the conditions, and therefore likely to be relevant to the disease or biological process under investigation. A suitably defined mathematical function (the t-statistic is a canonical choice) is used to assign a "relevance score" to each gene and a subset of genes selected on the basis of the scores. Here, the objects being selected are genes.

As a second example, consider the mining of sales records. The aim might be, for instance, to focus marketing efforts on a subset of customers, based on some property of their buying behavior. A suitably defined function would be used to score each customer by relevance, on the basis of his or her records. A set of customers with high relevance scores would then be selected as targets for marketing activity. In this example, the objects are customers.

Clearly, both tasks are similar; each can be thought of as comprising the assignment of a suitably defined relevance score to each object and the subsequent selection of a set of objects on the basis of the scores. The selection of objects thus requires the imposition of a threshold or cut-off on the relevance score, such that objects scoring higher than the threshold are returned as relevant. Consider the microarray example described above. Suppose the function used to rank genes is simply the difference between mean expression levels in the two classes. Then the question of setting a threshold amounts to asking how large a difference is sufficient to consider a gene relevant. Suppose we decide that a difference in means exceeding x is 'large enough': we would then consider each gene in turn, and select it as "relevant" if its relevance score equals or exceeds x. Now, an important point is that the data are random variables, so that if measurements were collected again from the same biological system, the actual values obtained for each gene might differ from those in the particular dataset being analyzed. As a consequence of this variability, there will be a real possibility of obtaining scores in excess of x from genes which are in fact *not* relevant.

In general terms, high scores which are simply due to chance (rather than the underlying relevance of the object) lead to the selection of irrelevant objects; errors of this kind are called false positives (or Type I errors). Conversely, a truly relevant object may have an unusually low score, leading to its omission from the final set of results. Errors of this kind are called false negatives (or Type II errors). Both types of error are associated with identifiable costs: false positives lead to wasted resources**,** and false negatives to missed opportunities. For example, in the market research context, false positives may lead to marketing material being targeted at the wrong customers; false negatives may lead to the omission of the "right" customers from the marketing campaign. Clearly, the rates of each kind of error are related to the threshold imposed on the relevance score: an excessively strict threshold will minimize false positives but produce many false negatives, while an overly lenient threshold will have the

opposite effect. Setting an *appropriate* threshold is therefore vital to controlling errors and associated costs.

Statistical hypothesis testing can be thought of as a framework within which the setting of thresholds can be addressed in a principled manner. The basic idea is to specify an acceptable false positive rate (i.e. an acceptable probability of Type I error) and then use probability theory to determine the precise threshold which corresponds to that specified error rate. A general discussion of hypothesis tests at an introductory level can be found in textbooks of statistics such as DeGroot and Schervish (2002), or Moore and McCabe (2002); the standard advanced reference on the topic is Lehmann (1997).

Now, let us assume for the moment that we have only one object to consider. The hypothesis that the object is irrelevant is called the null hypothesis (and denoted by H_0), and the hypothesis that it is relevant is called the alternative hypothesis (H_1). The aim of the hypothesis test is to make a decision regarding the relevance of the object, that is, a decision as to which hypothesis should be accepted. Suppose the relevance score for the object under consideration is t. A decision regarding the relevance of the object is then made as follows:

(1) Specify an acceptable level of Type I error p^*.
(2) Use the sampling distribution of the relevance score under the null hypothesis to compute a threshold score corresponding to p^*. Let this threshold score be denoted by c.
(3) If $t \geq c$, reject the null hypothesis and regard the object as relevant. If $t < c$, regard the object as irrelevant.

The specified error level p^* is called the significance level of the test and the corresponding threshold c the critical value.

Hypothesis testing can alternatively be thought of as a procedure by which relevance scores are converted into corresponding error probabilities. The null sampling distribution can be used to compute the probability p of making a Type I error if the threshold is set at exactly t, i.e. just low enough to select the given object. This then allows us to assert that the probability of obtaining a false positive if the given object is to be selected is at least p. This latter probability of Type I error is called a P-value. In contrast to relevance scores, P-values, being probabilities, have a clear interpretation. For instance, if we found that an object had a t-statistic value of 3 (say), it would be hard to tell whether the object should be regarded as relevant or not. However, if we found the corresponding P-value was 0.001, we would know that if the threshold were set just low enough to include the object, the false positive rate would be 1 in 1000, a fact that is far easier to interpret.

MAIN THRUST

We have seen that in the case of a single variable, relevance scores obtained from test statistics can be easily converted into error probabilities called P-values. However, practical data mining tasks, such as mining microarrays or consumer records, tend to be on a very large scale, with thousands, even millions of objects under consideration. Under these conditions of multiplicity, the conventional P-value described above no longer corresponds to the probability of obtaining a false positive.

An example will clarify this point. Consider once again the microarray analysis scenario, and assume that a suitable relevance scoring function has been chosen. Now, suppose we wish to set a threshold corresponding to a false positive rate of 0.05. Let the relevance score whose P-value is 0.05 be denoted by t_{05}. Then, in the case of a single variable/gene, if we were to set the threshold at t_{05}, the probability of obtaining a false positive would be 0.05. However, in the multi-gene setting, it is *each* of the thousands of genes under study that is effectively subjected to a hypothesis test with the specified error probability of 0.05. Thus, the chance of obtaining a false positive is no longer 0.05, but much higher. For instance, if each of 10000 genes were statistically independent, $(0.05 \times 10000) = 500$ genes would be mistakenly selected on average! In effect, the very threshold which implied a false positive rate of 0.05 for a single gene now leaves us with hundreds of false positives.

Multiple hypothesis testing procedures address the issue of multiplicity in hypothesis tests and provide a way of setting appropriate thresholds in multi-variable problems. The remainder of this Section describes two well-known multiple testing methods (the Bonferroni and False Discovery Rate methods), and discusses their advantages and disadvantages.

Table 1 summarizes the numbers of objects in various categories, and will prove useful in clarifying some of the concepts presented below. The total number of objects under consideration is m, of which m_1 are relevant and m_0 are irrelevant. A total of S objects are selected, of which S_1 are true positives and S_0 are false positives. We follow the convention that variables relating to irrelevant objects have the subscript "0" (to signify the null hypothesis) and those relating to relevant objects the subscript "1" (for the alternative hypothesis). Note also that fixed quantities (e.g. the total number of objects) are denoted by lower-case letters, while variable quantities (e.g. the number of objects selected) are denoted by upper-case letters.

The initial stage of a multi-variable analysis follows from our discussion of basic hypothesis testing and is