Outlier Detection by Rareness Assumption

Tomas Hrycej and Jochen Hipp

DaimlerChrysler AG, Research & Technology, Ulm, Germany

Tomas.Hrycej@daimlerchrysler.com

Abstract: A concept for identification of candidates for outliers is presented, with a focus on nominal variables. The database concerned is searched for rules that are *almost* universally valid, with rare exceptions. In statistical terms, for these rules, the hypothesis that the rule is universally valid except for random faults cannot be rejected. Outlier candidates are those values that violate these rules.

1 Introduction

Detecting faulty data items, or ,,outliers" in large data sets is a task of considerable interest. Of particular importance is the detection if the data are automatically processed. Since the outliers take on exceptional and unpredictable values, their effect may overshadow all genuine regularities in the data and prevent the algorithms from disclosing meaningful results.

On the other hand, it is virtually impossible to distinguish the outliers from the variations of the genuine data with ultimate certainty. An elegant way is with help of explicit stochastic models of correct data, or, in a complementary manner, explicit stochastic models of outliers. These models can then be analyzed, for example, in case of continuous variables by the methods of blind source separation (see e.g, [Ca00]). However, this approach is difficult to follow in the context of Data Mining tasks because of the lack of any prior information about the data.

Another challenge is the appropriate definition of the granularity at which numerical variables are analyzed and set into relationship with nominal variables. Theoretically, all information is captured in the joint probability distribution of all values. However, the joint probability distribution may be arbitrarily complex (in many cases, it is substantially different from Gaussian), and its approximation (e.g., by an expansion with help of high order moments) would be a computationally infeasible task. So it may be more appropriate to define discrete states of numerical variables and analyze their co-occurrence with the states of discrete variables by means of discrete statistics.

For data analysis, discarding ,,suspicious" data may lead to a bias in results of the analysis. However, in practice, it is usually not disastrous to discard a small fraction of data records. By contrast, working with unrecognized outliers may completely invalidate the results. So the usual procedure is to view the data and determine the data values that are candidates for

outliers. If their total number or proportion is not critical, they may be simply discarded. If the proportion of outliers is excessive, the whole data set may become useless unless the outliers are corrected.

Evaluation, discarding or correction of the outliers can also be done manually with help of expert understanding of the data. This step is relatively inexpensive if good candidates are available. By contrast, the candidate generation is a laborious task for which machine support is helpful. Such support is the main motivation for the work presented here.

2 Definition of atomic propositions

In this section, we define atomic propositions whose co-occurrence with others within the same record of a table may be of interest. For such atomic propositions, rules can be formulated, and outliers searched for.

Suppose there is a database table with a set of column variables with defined types. For simplicity, let us consider only the following types of variables: *char/varchar* (for nominal variables), *int* (for integer variables) and *float/double* (for floating point variables). While variables with few values constitute atomic proposition in a trivial way (by a variable having a particular value), the granularity on which real valued variables are investigated may (and must, for computational reasons) below the individual value level. So the discrete states to be analyzed can be defined, for example, as $var = value$ for nominal or integer variables with a limited number of distinct values (e.g., below 10), as well as $var < 0$, $var > 0$ and $var = 0$ for integer variables with many distinct values and floating point variables. Another state of interest may be the equality or nonequality of two numeric variables, formally $var1 < var2$, $var1 > var2$, $var1 = var2$.

Every such state represents an atomic proposition whose co-occurrence with others can be investigated (for example the co-occurrence of atomic proposition A defined as $color = red$ with B defined as $price > residual\_value$.

3 Outlier identification

As stated in Section 1, outliers can be separated from random data variations only statistically, with help of some prior knowledge or assumption about the correct data. For continuous variables, maximum entropy priors [O'94] can be assumed in the absence of any knowledge. Examples for such maximum entropy priors are Gaussian distribution for unbounded random variables with a given mean and variance, exponential distribution for positive random variables with a given mean and uniform distribution for random variables from a finite interval

For discrete states, the maximum entropy prior is the distribution with equal probabilities of all values. However, this assumption is so unrealistic that most variable from real data bases would violate it even without outliers.

This is why we decided to follow another approach. We will look for regularities that apply

at „almost all" cases with as few exceptions that it can be expected they are random faults. This amounts to the assumption that outliers are rare, thence the „rareness assumption" in the title of this work. Of course, in this way, we can find only the *candidates* for outliers since they cannot be safely distinguished from very scarce, but correct occurrences.

The simplest case is that of a single atomic proposition. Suppose an atomic proposition A, of some of the types given in Section 2, were universally valid in a „clean" data base. Suppose further, in a real, partially erroneous database, this proposition is violated with probability p. For example, the variable $price$ would be nonnegative in a clean database (A corresponding to $price \geq 0$), but be erroneously assigned a negative value with probability p.

Then the occurrences of the fault $\neg A$ are binomially distributed, and the probability that $\neg A$ occurs in $k_{\neg A}$ out of n data records is

$$\binom{n}{k_{\neg A}} p_{\neg A}^{k} (1-p)^{n-k_{\neg A}} \tag{1}$$

If the number of occurrences is $k_{\neg A}$ and the cumulative probability

$$P(k \geq k_{\neg A}|n) = \sum_{i=k_{\neg A}}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1 - \sum_{i=0}^{k_{\neg A}-1} \binom{n}{i} p^i (1-p)^{n-i} \tag{2}$$

is below the given significance level α, then the hypothesis that this number of occurrences of $\neg A$ can be explained by the fault process of the described type can be rejected on this significance level [SOA99]. Then, we can assume that A is not universally valid, and $\neg A$ may be correct values.

A widespread (and probably computationally the fastest) hypothesis test for the binomially distributed number of occurrences consists of an approximation by the binomial distribution by the Gaussian with the same mean and standard deviation, that is, $N(np, \sqrt{np(1-p)})$. The number of occurrences $k_{\neg A}$ would then be compared with $np + q_\alpha \sqrt{np(1-p)}$ with q_α being the α-quantile of the standard normal distribution. However, this approximation is justified only for relatively high values of p, which can hardly be expected in our case where p is the probability of a single data fault. So it is advisable to perform the test with help of the original formula (2). An efficient way of computation is with help of the recursive formula

$$\begin{array}{lcl} P(n,0) & = & (1-p)^n \\ P(n,k) & = & P(n,k-1)\frac{n-k+1}{k}\frac{p}{1-p} \end{array} \tag{3}$$

If the hypothesis cannot be rejected, we still cannot confirm the opposite, that is, that the occurrences of $\neg A$ are faults. However, they are obviously justified candidates for faults. To confirm this hypothesis by statistical means, we would have to evaluate the probability β of the fault of the second kind (accepting an invalid hypothesis), which is substantially more complex, and requires further assumptions [SOA99]. But the generating justified candidates is sufficient for our aim.

We can proceed by looking for nonatomic expressions such as rules of the form $X \rightarrow A$ with X being a logical expression made of atomic propositions. In practice, X will typically consist of a single atomic proposition or a conjunction of two to three propositions. The co-occurrences of X and A are characterized by the quadruple

$$(n, n_X, n_A, n_{X \wedge A}) \tag{4}$$

The cumulative probability is then received by substituting n_X for n and $k_{X \wedge \neg A}$ for $k_{\neg A}$ in (2), and tested for excess of α.

So every rule can be, with help of its corresponding tuple (4), determined to belong to one of the following classes:

- (A) those that are universally valid, recognized by equality $n_X = n_{X \wedge A}$,
- (B) those for which the hypothesis of universal validity can be rejected, that is, those for which (2) is below the significance level α, and
- (C) those that are not universally valid, but whose hypothesis of universal validity cannot be rejected, that is, (2) exceeds the significance level α; in other words, it is not inconsistent to assume that the rule is violated only be faulty records, or outliers

The procedure for outlier detection consists of the following steps

1. Generating rules (i.e., atomic or nonatomic expressions)
2. Assigning the rules to one of the above three classes
3. Retaining the rules of class C

The values of fault probability p and significance level α are to be a priori specified. While for significance level some of the standard values such as 0.01 or 0.001 can be taken, the fault probability has to be deliberately set. Fortunately, its effect is monotone - a low p will produce less outlier candidates, a high p more candidates.

4 Computing considerations, experience, and future work

The presented concept has been implemented as a PERL script with an access to a MySql-database. The atomic propositions have the form of character strings which, in turn, can be directly used for the construction of SQL queries to receive the counts (4).

The process of rule generation is computationally intensive. For N_n nominal variable having on average M distinct values and N_r numeric variables, there are $P = N_n M + 3N_r + 3N_r(N_r - 1)/2 = N_n M + 1.5N_r + 1.5N_r^2$ atomic propositions. So there are $P(P-1)/2$ simple rules of the mentioned type, which can easily amount to several millions (e.g., for $N_r = 50$, the number of rules exceeds seven millions). For search of larger rules, frequent itemset algorithms are efficient, as long as minimum support (i.e., lower bound for $N_{X \wedge A}$) can be specified. (Specifying the minimum support may be a serious obstacle since the

„appropriate" minimum support is difficult to determine. By contrast to, for example, the significance level or fault probability, which hardly depend on the domains of the variables, the support for a atomic proposition of the type $< variable >=< value >$ decreases with increasing number of distinct values of $< variable >$.)

A typical mode of use of the algorithm consists of (1) setting up the table to be analyzed (possibly as a join of several tables), (2) determining the rules violated by potential outliers and (3) listing the records violating the rules.

The algorithm has been applied to several real data bases, containing data sets from the fields of quality insurance and fleet car management. The parameters of the algorithm have been set to $p = 0.001$ and $\alpha = 0.001$. The following example illustrates the output: all rules are shown that concern the variable *Migr* and that are violated by some outliers. (The value *'Yes'* of variable *Migr* identifies the records which were migrated from a previous system and are thus particularly error-prone.)

ValR=0	→	Migr='Yes',	1 exception
IntR<0	→	Migr='Yes',	1 exception
CapR<0	→	Migr='Yes',	1 exception
Migr=NULL	→	NOT(StatVRw='Stock')	6 exceptions

For example, the first rule says that all records in which the variable *ValR* is zero are migrated records, with a single exception, which is suspect of being an outlier.

Viewing the outlier records („exceptions") enabled their correction or removal from the data base in order to improve the quality of the subsequent Data Mining analysis.

Our next step will be the thorough evaluation of the basic approach on several data bases from different domains. Furthermore we plan to investigate whether the results improve by allowing rules to contain a conjunction of two or even more atomic propositions in the antecedent. Due this extension implying increasing complexity we plan to address the problem of efficient rule generation by means of frequent itemset generation, e.g.[AS94, HGN00]. Although we are aware of the fact that the (necessary) introduction of minimum support restricts the set of rules discovered we expect that this may bring about valuable results.

Literatur

[AS94] Agrawal, R. und Srikant, R.: Fast algorithms for mining association rules. In: *Proceedings of the 20th International Conference on Very Large Databases (VLDB '94)*. S. 487–499. Santiago, Chile. June 1994.

[Ca00] Cardoso, J.-F.: Entropic contrasts for source separation: geometry and stability. In: Haykin, S. (Hrsg.), *Unsupervised adaptive filtering, Volume I*. Wiley, New York. 2000.

[HGN00] Hipp, J., Güntzer, U., und Nakhaeizadeh, G.: Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations*. 2(1):58–64. July 2000.

[O'94] O'Hagan, A.: *Kendall's Advanced Theory of Statistics, Volume II B: Bayesian Inference*. Arnold. London. 1994.

[SOA99] Stuart, A., Ord, K., und Arnold, S.: *Kendall's Advanced Theory of Statistics, Volume II A: Classical Inference and the Linear Model*. Arnold. London. 1999.