JMLR: Workshop and Conference Proceedings 17 (2011) 56–62 2nd Workshop on Applications of Pattern Analysis

A Software System for the Microbial Source Tracking Problem

David Sánchez DAVID.SANCHEZ-MENDOZA@EST.FIB.UPC.EDU
School of Computer Science, Technical University of Catalonia, Barcelona, Spain

Lluís A. Belanche BELANCHE@LSI.UPC.EDU
Department of Software, Technical University of Catalonia, Barcelona, Spain

Anicet R. Blanch ABLANCH@UB.EDU
Department of Microbiology, University of Barcelona, Barcelona, Spain

Editor: Tom Diethe, José L. Balcázar, John Shawe-Taylor, and Cristina Tîrnăucă

Abstract

The aim of this paper is to report the achievement of ICHNAEA, a fully computer-based prediction system that is able to make fairly accurate predictions for Microbial Source Tracking studies. The system accepts examples showing different concentration levels, uses indicators (variables) with different environmental persistence, and can be applied at different geographical or climatic areas. We describe the inner workings of the system and report on the specific problems and challenges arisen from the machine learning point of view and how they have been addressed.

Keywords: Microbial Source Tracking, Real-World Machine Learning Applications

1. Introduction

Microbial source tracking (MST) is a recently coined term that includes different methodological approaches that pursuit the determination of the origin of fecal pollution in water by the use of microbial or chemical indicators (**?**). Nowadays, fecal pollution in water is one of the main causes of health problems in the world, and is associated with several thousands of deaths per day, being a main vehicle of pathogen transmission. The problem is important from both the scientific and legal points of view. Particularly interesting are those prediction models that use a minimum number of variables, given the large technical and monetary costs that are implicit in the collection of the source data matrix and in obtaining new examples.

There is a clear trend in MST studies to define specific indicators for different fecal sources and to establish standardized methodologies by an easy routine application for the enumeration of these indicators. Most of the research carried out has been focused on defining new indicators and suitable methodologies for detection and enumeration. Most of the many studies that develop predictive models for MST have been based on the definition of predictive models at the point of source, assessing mainly (but not exclusively) the specificity and sensitivity of indicators and/or their combinations. Progressively, the need has arisen to assess the effects of the *dilution* of the contributions of fecal pollution in receiving waters as well as the *persistence* of these indicators on the environment.

Machine learning (ML) is a discipline concerned with the design and study of algorithms that allow machines (computers) to incorporate models of phenomena based on empirical data about the modelled phenomena. The simplest instance of the MST problem (differentiating human origin from non-human origin in recently and heavily polluted waters) has been already solved to satisfaction using ML techniques (**?**).

There still are many problems to face, including independence of geographical location (or at least portability under similar watershed conditions), nature of the dominant fecal pollution contributions (anthropogenic or non-anthropogenic), persistence of indicators and their measured parameters, effects of dilution in watersheds and presence of complex mixtures from several distinct animal species. The final step would be the development of predictive models to be applied in water management and decision making processes.

In this paper we report the development of Ichnaea[1], a software system developed entirely in R (**?**) that represents a step forward in the solution of the general MST problem. The system accepts examples showing different fecal concentration levels (dilutions), uses indicators with different environmental persistence, and can be applied at different geographical or climatic areas, thus tackling many of the above open problems. To reach this achievement, a number of problems have been encountered and addressed, which may be of interest to a broad audience within the ML community.

The rest of the document is organized as follows. In Section 2 the MST problem is described in more detail. In Sections 3 and 4 we report on the specific problems and challenges arisen from the ML viewpoint and how they have been addressed. In Section 5 the inner workings of the system are described and preliminary performance results are reported. We close with some conclusions and a brief description of the remaining work.

2. Microbial source tracking

The determination of the source of fecal contamination in water is necessary to estimate health risks associated to polluted waters, to take measures to clean polluted waterways and to solve legal problems concerning who should clean the polluted water bodies. At present, there is a widespread agreement that no single microbial or chemical indicator is able to best determine sources of fecal pollution. In particular, it has been reported that at least two parameters are needed to accurately differentiate between two distinct fecal pollution sources: one specific indicator that identifies the source and another, universal indicator that provides information on the fecal load (**?**). In consequence, the indicators should be carefully selected based on appropriate statistical analyses.

When the *age* of fecal pollution is also an issue, identification becomes more complicated and thus the availability of reliable decay rates is of primary importance to MST (**?**). In surface waters impacted by runoff at point sources, the use of ratios between parameters can help in determining the origin of fecal contamination. Nonetheless, persistence studies are needed to provide complementary information, addressing the effects of environmental aspects like temperature, solar radiation, salinity, pH, chemical pollutants, water filtration, turbidity, starvation, predation and presence of heavy metals, among others. Other relevant issues are the use of methods based either on genomic targets or in quantification of new microbial tracers through library-independent methods, the combination of methods, and

1. Ichnaea (or $I\chi\nu\alpha\iota\eta$) was the *tracing* goddess, one of the female Titanes.

comparative and integrated studies between research groups with standardised procedures to avoid differences in implementation.

Most of the modeling approaches used in MST make use of traditional techniques, such as discriminant analysis or nearest neighbours, and a few use artificial neural networks, such as **?**. The need has arisen for new approaches merging take the best from several disciplines (microbiology, epidemiology, ecology, statistics and computer science, among others). These approaches should account for the limitations already found in the development of predictive models (**?**).

3. Problems and Challenges

The starting point of the modelling is a data matrix consisting of 103 examples described by 26 indicators (microbial and chemical) with no missing values. These examples are labelled in two classes (the known fecal source origins): 54 human origin and 49 non-human origin (pigs, poultry and cows, mainly). The straight solution would seem to select a set of off-the-shelf (parameterized) classifiers, split the data in learning and testing parts, fit the classifiers on the learning part using a resampling technique like cross-validation and choose that combination classifier/parameters that yields the lowest cross-validated error. Assessment of performance could be obtained by evaluating the model on the testing set.

This approach is not going to work. There are at least three technicalities serious enough to hinder or, in the present case, even prevent a standard ML solution:

1. The examples in the data matrix are at *point source*: they were taken right in the spot of contamination. In practice, this will not be the case and we say that the example is *diluted*. A dilution factor of 25 (say), represents that the theoretical value is divided by 25. Moreover, there comes a point where the concentration of a measured variable may fall below its detection threshold. In this case, the value present in the example is constant and equal to the detection threshold (a very small value).

   The problem, from the standpoint of ML, is twofold. First, relative class morphology is altered, hence models that are good for data at point source may not be as good for increasing dilution factors. Second, there comes a point in which the variable simply disappears (indicated by the mentioned detection threshold).

2. The examples in the data matrix are expressed at *zero-time*: they were taken right after contamination took place. In practice, this will not happen and the example to be predicted will be *aged*. Moreover, the distinct variables age following different processes that are not completely understood. These processes are altered depending on season conditions, as mentioned above. The aging process is merged with the dilution factor and alters even more the relative shape of the two classes, specially considering that each variable evolves differently.

3. An important characteristic of the system is that it is based on the data matrix supplied by the user (thus different users can enter different matrices). The data matrix is intended to reflect the information the user has been able to collect about his/her MST study, expressed in the form of specific biological or chemical tracers. This matrix should be regarded as *maximal* and a strong need arises to reduce the

amount of information needed for the prediction. The immediate consequence is that end-users will supply only a fraction of the variables in the matrix, depending on varying technical, geographical or monetary conditions.

From the standpoint of ML, this is a serious concern. Although feature selection can be conducted to reduce the number of variables to a minimum while retaining discriminating ability, the reduced subsets will depend on dilution factor and age. Moreover, there is no guarantee that users will be able to supply all or part of the obtained relevant subsets.

In short, the aim of the system is to distinguish the origin of fecal pollution out of water examples that have not been taken at point source (viz. *diluted*) and/or present some delay between dumping and measurement time (viz. *aged*). It is very important to mention that both the dilution factor and the aging time are completely unknown.

4. Some Solutions to the Problems and Challenges

The first action to be taken is to estimate the time passed between dumping and measurement. Field studies have been recently performed to evaluate persistence in the environment at different seasons (**?**). These works report a set of empirical measurements at different stipulated times for many of the indicators used in the present study. Based on several realizations of these assays, it was found that the hypothesis of log-linear relations between the measured value and time was tenable. The die-offs of the involved indicators have been calculated and integrated on the prediction system[2]. These measures were taken at the point of source (and thus are non-diluted).

To estimate the elapsed time since disposal in the receiving watershed, two additional pieces of information reveal as key facts:

1. The dilution and aging processes are, in practical terms, independent: the die-off (persistence evolution) of an indicator does not depend on its concentration level.

2. The persistence is a function of the indicator and the environment and thus is not dependent on the source (i.e., persistence of the same indicator in equal conditions is the same for humans and non-humans).

Consider $S = \{(x_1, \log_{10}(y_1)), \ldots, (x_n, \log_{10}(y_n))\}$ a univariate data sample. Let $f(x) = ax + b$ represent the regression line on S, obtained by ordinary least squares (OLS). Let $S_\alpha = \{(x_1, \log_{10}(y_1/\alpha)), \ldots, (x_n, \log_{10}(y_n/\alpha))\}$ a dilluted data sample, where $\alpha \geq 1$ is the (unkown) dilution factor. Then the regression line obtained by OLS on S_α would be $f_\alpha(x) = ax + b - \log_{10}(\alpha)$. Consider now a variable V_i with known measured value v_i in a given example to be predicted. The theoretical equation for V_i is $\log_{10}(\frac{\widetilde{v}_i}{\alpha}) + a_i t = v_i$, where a_i is the obtained slope and $\widetilde{v}_i$ is the unknown non-diluted zero-time measurement.

In prediction time, we are given a collection of available variables $V_1, \ldots, V_N$ with measured above-threshold values $v_1, \ldots, v_N$. The evaluation of these equations need the value of α. However, if we subtract two of the equations we arrive at:

2. It should be noted that two sets of assays were carried out: one for winter and another for summer.

$$(a_i - a_j)t + \log_{10}(\widetilde{v}_i) - \log_{10}(\widetilde{v}_j) = v_i - v_j \quad (1)$$

i.e. the *difference* in time behaviour of a pair of variables is not affected by the dilution process (at least not until one of the variables reaches its threshold). The quantities $\log_{10}(\widetilde{v}_i), \log_{10}(\widetilde{v}_j)$ are not directly available but, recalling that the supplied data matrix consists of non-diluted zero-time measurements, a set of $N(N-1)/2$ equations of the form (Equation (1)), can be obtained by replacing every *pair* $\log_{10}(\widetilde{v}_i), \log_{10}(\widetilde{v}_j)$ with the corresponding values in the data matrix. Then an estimation for the elapsed time t^* can be obtained by OLS in this system of equations. Once this is known, an estimation for the dilution factor α^* can be obtained using any variable and the relation $\log_{10}(\alpha^*) = t^* a_i + b_i - v_i$. Reversing time, $\widehat{v}_i$ is the *deaged* value of the (still dilluted) variable V_i: $\widehat{v}_i = v_i - a_i t^*$, $1 \leq i \leq N$.

To tackle the problem that only a *varying* fraction of the variables in the matrix will be supplied for a new example to be predicted, the idea is to *recycle* the data matrix by using it in a set of independent training processes for different values of the dilution factor, as follows. First, a set of equidistant dilution factors is selected in the interval $[1, 500]$. For each generated subinterval, the data matrix is diluted to that factor and different models are developed using this diluted matrix. Model selection is limited to those requiring 2 to 3 indicators, for the reasons explained in Section 2. The result is a set of diverse and simple models trained to respond to different ranges of dilution, that are used in a second step to test their suitability and accuracy to predict new water examples with fecal pollution.

5. Overall System Description and Validation

The starting point for the system is a data matrix of the set of MST indicators selected by the user and the obtained values for those indicators by taking different fecal polluted water examples at point source from the geographical and climatic areas where the study is performed. The main characteristic of the matrix is that all water examples are taken at point source (heavy fecal pollution) and with no time delay between dumping and sampling (zero time). Given the matrix, the system pre-computes a number of predictive models (may be several hundreds) for different sampled dilution factors, using a set of off-the-shelf two-class classifiers, such as linear discriminant analysis, the support vector machine, nearest neighbours and logistic regression (**?**). The suitability and accuracy of each model is assessed with cross-validation. Given a new example, defined by certain user-chosen indicators (that must be a subset of those present in the initially entered data matrix), the relations among the obtained values for the above-threshold supplied indicators are used to calculate the estimation t^* of the time passed since the pollution took place as well as the estimated dilution factor α^*. These values are then used to *deage* the example as shown in Section 4: the measured values of the example are adjusted using the available reduction curves to obtain its corresponding values at zero time.

At this moment, a *feasible* subset of models is selected among the set of models previously computed. This feasible set is given by those models matching the estimated dilution factor that cover the example to be predicted. We say that a model *covers* an example if every variable needed in the model is present in the example.

Once all the selected models make their prediction, an overall prediction is determined by a simple majority vote. The output delivered to the end user includes the prediction of

fecal source for the example, a measure of the system's confidence on this prediction (given simply as the degree of agreement of the overall prediction), the values t^* and α^*, and a recommendation of certain additional indicators which, if measured by the user, could increase the confidence of the prediction if requested. This last information is computed analysing the estimated performance of all models belonging to the estimated dilution factor that did not cover the example to be predicted. It is expected that the interaction system-user is not reduced to a single query. Given the output of the first query about an example, the user may decide (or not) to invest further efforts in acquiring additional indicators.

The described approach has been validated using the data matrix from a previous international project (**?**). The described computer-based integrated prediction system showed an estimated performance of $75 - 80\%$ correct classification on examples aged from 0 to 150 hours, that can be diluted by a factor of up to 500 (Fig. 1). The prediction accuracy depends of the number and composition of the measured indicators and the true dilution degree and persistence of the example. As could be expected, less diluted and freshly polluted examples lead in general to a better accuracy.

These results should be considered promising, specially considering that the majority class has probability 52.4%, and the great number of approximations and estimations that the system is forced to compute (among others, the log-linear decays are only average behaviours; moreover, although we work with summer/winter sets of measurements, certainly environmental conditions in February may be quite different than in November). In this vein, although the identification of the origins of fecal pollution in watersheds has been an intensive research field in recent years, the ability to make precise and significant statements about the behavior of fecal polluted environmental waters decreases as the complexity of the scenario increases, thus setting a limit on practical predictive performance.

6. Conclusions

Ichnaea, a prototype computer-based integrated system for predictions on Microbial Source Tracking (MST) is presented. The system can be trained by users with their own data matrix developed with the MST indicators that they select within their geographical and climatic environment. The system provides the accuracy and the precision of the prediction, the estimated degree of dilution and the age of the pollution for the analysed example. Moreover, complementary MST indicators (among those constituting the user-supplied data matrix) are suggested to further improve on the confidence of the MST prediction.

Further research includes analysing the presence of several distinct animal species, thus making the problem a multiclass classification task, with all the associated difficulties. An important avenue of further work is found in providing posterior probabilities for each class (animal type) in places where pollution comes from a mixture of sources. For example, a sample may show 80% of fecal pollution of human origin and a 20% of pig origin, a situation that better represents reality in many practical occasions.

Acknowledgments

This study was funded by the Spanish Government project CGL2007-65980-C02-01.

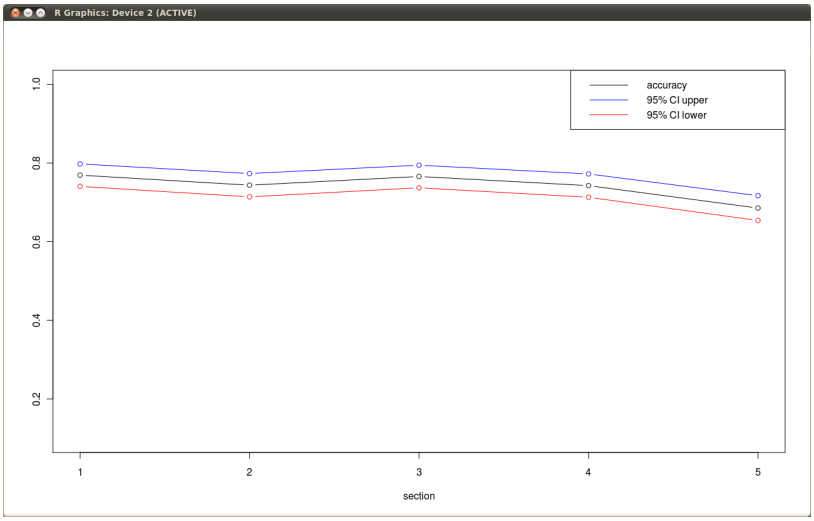

Figure 1: Accuracy of the predictions. The x-axis is the (true) dilution degree of the predicted example, divided by 100; the y-axis is the overall obtained accuracy. The 95% upper and lower confidence bands are added to the mean accuracy line.

References