Evolutionary Mining for Multivariate Associations in Large Time-Varying data sets: a Healthcare Network Application

Narine Manukyan
University of Vermont (UVM)
Burlington, VT, USA
Narine.Manukyan@uvm.edu

Margaret J. Eppstein
University of Vermont (UVM)
Burlington, VT, USA
Maggie.Eppstein@uvm.edu

Jeffrey D. Horbar
Vermont Oxford Network,UVM
Burlington, VT, USA
Jeffrey.Horbar@uvm.edu

ABSTRACT

We introduce a new method for exploratory analysis of large data sets with time-varying features, where the aim is to automatically discover novel relationships between features (over some time period) that are predictive of any of a number of time-varying outcomes (over some other time period). Using a genetic algorithm, we co-evolve (i) a subset of predictive features, (ii) which attribute will be predicted (iii) the time period over which to assess the predictive features, and (iv) the time period over which to assess the predicted attribute. After validating the method on 15 synthetic test problems, we used the approach for exploratory analysis of a large healthcare network data set. We discovered a strong association, with 100% sensitivity, between hospital participation in multi-institutional quality improvement collaboratives during or before 2002, and changes in the risk-adjusted rates of mortality and morbidity observed after a 1-2 year lag. The results provide indirect evidence that these quality improvement collaboratives may have had the desired effect of improving health care practices at participating hospitals. The proposed approach is a potentially powerful and general tool for exploratory analysis of a wide range of time-series data sets.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*knowledge acquisition*; I.5.m [**Computing Methodologies**]: Pattern Recognition—*Miscellaneous*

General Terms

Algorithms, Experimentation, Design

Keywords

Genetic algorithms, exploratory multivariate data analysis, time series.

1. INTRODUCTION

The rapid growth of technology has facilitated widespread collection and storage of vast amounts of time-varying data. This data undoubtedly contains a wealth of potentially valuable information regarding relationships between various time-varying features and outcomes. However, the very size of these databases is an impediment to knowledge discovery, creating a need for automated exploratory analysis tools [2]. The challenge of exploratory data analysis is that one should not only identify features that are associated in a potentially non-linear manner, but also determine which outcome(s) those features are associated with and time dependent aspects of the association. In this paper we develop a general tool that uses genetic algorithms and classifiers to find novel multivariate associations between features in time varying data. We first validate the approach using synthetic data and then apply it to a real world problem of finding potential associations between 18 different patient outcomes over a 10 year period and different hospital collaborations in the Vermont Oxford Network (VON), a worldwide network of neonatal intensive care units designated to promote dissemination of effective healthcare practices. This data set is difficult to analyze because it includes a large number of time-varying patient outcomes as well as several different types of time-varying interactions between member hospitals. It is not clear which interactions (if any) may be associated with which changes in patient outcomes or, if such an association exists, over which time frames the association is strongest. Genetic algorithms (GAs) are known to be effective for feature selection [1], but we are not aware of any research that does feature selection and predicted attribute selection for time series data.The unique contribution of this study is to apply a GA for simultaneous selection of features, feature time frames, which attribute to predict, and over what time period to predict it.

GECCO'12, July 7-11, 2012, Philadelphia, Pennsylvania, USA.
ACM 978-1-4503-1177-9/12/07.

2. METHODS

We use a GA to simultaneously estimate four important aspects of multivariate time-series analysis: (i) a subset of features to be used as input into some sort of statistical predictor, (ii) which attribute we can best predict from these features, (iii) a dividing year that partitions the time-series, and (iv) a time lag to be added to the dividing year to indicate a possible delay in outcome change. Fitness is determined by measuring how well the values of the selected features before the dividing year can be used to predict changes in the selected attribute before the dividing year and after the dividing year + lag. For brevity, we refer to this method as GAMET (Genetic Algorithm for Multivariate Exploration of Time-varying data). For feature selection, we are using binary flags that indicate whether the given feature is included in the final features subset or not. To evolve the time series component we evolve the dividing year and lag, both of which are represented as gray-coded integers

in the chromosome. Finally, we evolve a gray-coded index specifying which single attribute (from a list of potentially predicted attributes) is to be predicted. To calculate the fitness of an individual, we first process the data for the included features, using the dividing year and lag as described above. We then pass these time-processed features as inputs to the classifier, and compare the predicted classes to class outcomes of the attribute specified by the predicted attribute index, averaged prior to the dividing year. The data is divided into training and testing sets, using a parameter to control the percentage of the data used for training (80% for our experiments). We use Latin hypercube sampling to ensure adequate distribution of samples in the training and testing sets for highly unbalanced classification problems. After the training phase we evaluate the classifier performance using the following formula:

$$fitness = \frac{FP}{(FP+TN)} + \frac{FN}{(FN+TP)} + \frac{(FP+FN)}{2(TP+FP+TN+TP)}$$

where FP is the number of false positives, TP is the number of true positives, FN is the number of false negatives and TN is the number of true negatives. The first two terms represent the proportion of samples in each class that were classified incorrectly, whereas the last term is the proportion of the overall misclassified samples (first two terms are helpful for unbalanced classes). We employ two different classifiers in this paper. For the synthetically generated data set, we were able to use a naïve Bayes quadratic discriminant analysis (DA) classifier. However, because the VON data set violated so many assumptions of the DA (e.g. normality), for this application we used a non-parameter counterpropogation artificial neural network (CPNN) classifier.

3. DATA AND EXPERIMENTAL DESIGN

To test GAMET we used synthetic data with known associations between a subset of feature vectors and one of the attributes to predict. The level of added random noise was increased as the dividing year and lag got farther from the target values, as one might expect to see in real time series data. We ran 10 replicates of the GA for each of 15 synthetic problems (using 2 , 5 and 8 true variables out of 100 total variables). For our real world dataset we consider four VON sponsored interactions among hospitals: (i) participation in VON annual meetings; (ii) preparation of case studies that were presented at VON meetings; (iii) participation in VON-sponsored team collaboratives, which are 2-year long team projects where multidisciplinary quality improvement teams from participating hospitals work together to identify and implement potentially better health practices, and (iv) co-authorship on publications resulting from VON-related activities. We then use 18 risk-adjusted measures over the period of 2001 through 2010, representing the health outcomes of over half a million very low birth weight infants to see if we can classify individual hospitals as participants or non-participants in any of the 4 types of VON sponsored interactions. We ran 10 replicates of the GA using the CPNN-based fitness function. Because both the DA and the CPNN can still classify well even with a certain number of excess features given as inputs, we subsequently intersected the feature sets of the best individuals resulting from each of the 10 replicates.

4. RESULTS AND DISCUSSION

In all 10 replications of each of the fifteen 100-feature synthetic problems GAMET was able to correctly identify the dividing year, lag, which attribute to predict (labeled "output"), and all of the 2, 3, or 8 true features, using the DA-based fitness function. As the number of true features increased, the tendency of GAMET to return excess features also increased, but intersecting the selected feature sets largely removed this problem. On the VON data set, all 10 runs consistently returned a dividing year of 2002, and discovered that participation in VON-sponsored team collaboratives was the attribute that could most accurately be classified. In 7 of the 10 runs, the lag was determined to be 2 years, whereas in the remaining 3 runs the lag was determined to be 1 years. From a clinical standpoint, these results are encouraging, because the team collaboratives sponsored by the VON up through 2002 had included teams focussing on outcomes selected by the GA (e.g., intravectricular hemorrhage, chronic lung disease). The health outcome features selected as input to the CPNN-based fitness function were also relatively consistent between the 10 runs. In all cases, the CPNN was able to predict the "true positives" in the smaller class (participants) with 100% accuracy (i.e., based on the selected health outcomes, the CPNN could correctly predict which hospitals *had* participated in a VON-sponsored team collaborative during or before the dividing year). However, the classifier was not able to use the selected health outcomes to accurately predict the "true negatives" (hospitals that didn't participate in any VON-sponsored team collaboratives during or before the dividing year). This result is actually to be expected, because there are many influences in healthcare practices at VON member hospitals that are independent of participation in VON-sponsored activities (and are consequently not in our database) that are expected to contribute to changes in health outcomes. Although the proposed method was originally developed for analysis of the VON healthcare network data set described here, the GAMET approach is a potentially powerful and general tool for exploratory analysis of a wide range of time-series data sets.

5. ACKNOWLEDGMENTS

This work was funded in part by the NIH Eunice Kennedy Shriver National Institute of Child Health & Human Development award 1R21HD068296.

6. ADDITIONAL AUTHORS

Kathleen A. Leahy, Michael J. Kenny (University of Vermont & Vermont Oxford Network, Burlington, VT, USA), Shreya Mukherjee, and Donna M. Rizzo (University of Vermont, Burlington, VT, USA).

7. REFERENCES

[1] M. ElAlami. A filter model for feature subset selection based on genetic algorithm. *Knowledge-Based Systems*, 22(5):356–362, 2009.

[2] D. Reshef, Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.