Applications of Spatio-Temporal Data Mining to North Platte River Reservoirs

Abhinaya Mohan
Department of Computer Science
University of Nebraska, Lincoln
269-779-7255
amohan@cse.unl.edu

Peter Z. Revesz
Department of Computer Science
University of Nebraska, Lincoln
571-201-5639
revesz@cse.unl.edu

ABSTRACT

We propose a spatio-temporal data mining method based on support vector machines regression and spatio-temporal feature reduction by principal component analysis. We apply the spatio-temporal data mining method to derive an automated controller for the reservoirs of the North Platte River. The automated controller opens and closes dams to efficiently and accurately control the reservoirs' water levels.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *data mining, spatial databases and GIS.* I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – industrial automation.

General Terms

Algorithms, Performance, Experimentation.

Keywords

Spatio-temporal data mining, regression, water reservoir, classifiers and prediction.

1. INTRODUCTION

Spatio-temporal data occur in a variety of forms [3,6,8,12] either as a series of discrete snapshots or as a continuous representation that is obtained by some interpolation method [9]. Spatio-temporal objects include moving objects [1], epidemic regions [16], and numerous changing geographic features.

River reservoir systems are complex spatio-temporal objects. River reservoir systems contain a number of dams that are opened or closed by human operators in order to lower or raise the reservoirs' water levels. Even for a single reservoir, numerous interactions between the input and the output parameters must be considered to determine the optimal release value. For multiple reservoirs, the storage value dictated of a downstream reservoir is dependent on the release levels of the upstream reservoirs.

River reservoir system spatio-temporal data mining aims to derive an expert system or an automated controller for the reservoirs' operations. Most data mining employ some linear classifiers [14] and integrate data from different sources [13,15], but to provide a good fit, a spatio-temporal data model needs to capture the non-linear relationship among the spatial and temporal variables.

Earlier spatio-temporal data mining proposals include, Cao et al.'s [2] data miner for spatio-temporal events, Dong's [5] wavelet transformations to develop efficient non-linear models, Mennis and Liu's [10] association rule mining by concept hierarchies, Ohashi and Torgo's [11] wind speed forecaster, and Tsoukatos and Gunopulos' [18] DFS_MINE algorithm for identifying frequent spatio-temporal patterns in environmental data. Several hydrological optimization works employ artificial neural networks (ANNs). For example, Coulibaly et al. [4] present a dynamic ANN model to optimize reservoir inflow values. Glezakos et al. [7] implement a neural networks model on time series dataset created using evolutionary clustering. The heuristics learned during meta-learning are included in the training phase of the model to develop models for watershed management. Solomatine et al. [17] develop hydrodynamic models for reservoir and ground water control using ANN-based optimization.

In this paper, we propose a new spatio-temporal data mining method, which combines support vector machines with regression analysis and principal component analysis feature selection. We apply the new spatio-temporal data mining method to the North Platte River reservoir system.

2. SPATIO-TEMPORAL DATA MINING

For the spatial model, the support vector machine (SVM)-based classifier takes as input the current (time *t*) values of the variables for each of the reservoirs. For the spatio-temporal data model, the values of the variables at previous time instances (*t-1, t-2* and *t-3*) are also considered in addition to their current values. The SVM kernel function is the major bottleneck for performance. Our implementation of the SVM with regression employs a quadratic kernel function of the form:

$$K(x,y) = (x^T y + C)^2 \qquad (1)$$

Data is usually split into two data sets: (a) training and (b) testing. The parameters of the data mining are estimated through the training phase. The estimated parameters are evaluated in the testing phase. The testing data must not contain patterns from the training dataset. The parameters of the data mining are updated until the testing phase maps the patterns to above the minimum acceptable performance limit. Including temporal information into the spatial data drastically increases the dimensions of the aggregated spatio-temporal data. The resulting situation is that there is a relatively large number of attributes but only no more

IDEAS'14, July 07-09, 2014, Porto, Portugal
Copyright ©2014 ACM 978-1-4503-2627-8/14/-7 $15.00.
http://dx.doi.org/10.1145/2628194.2628221

than hundreds of instances of the dataset. However, all of the input features do not necessarily have the same degree of descriptiveness. The target output feature might have a stronger degree of correlation with a definite set of the input features. Hence we employ a principal component analysis (PCA) for dimension reduction. Figure 1 shows the pseudo code.

```
Read(Data);
A = transpose(Data);
[n, m] = size(A);
Amean = mean(A);
AStd = standard deviation(A);
--Covariance matrix of input database--
V = covariance(A);
B = calculate zscore(A);

-- PCA from covariance--
[coeff,score,latent] = princomp(B);
-- Principal components analysis--
PC = coeff;

---Explained variance--
explainedvar=cumsum(variance(score))/sum(variance(score);
Input variance value k
J = length(k)
for i = 1 to J
     if explainedvar(i) <= K
        count = i;
    end
end
---PCA component selection—
Selected PC components=PC(:,1:count);
Z1=(((B *selected_PC ) *  selected_PC')* epmat(AStd,[n,1]))
       + repmat(Amean,[n, 1]);
Data_set=transpose(Z1);
return explainedvar(count);
return number of components;
```

Figure 1. The spatio-temporal data mining algorithm.

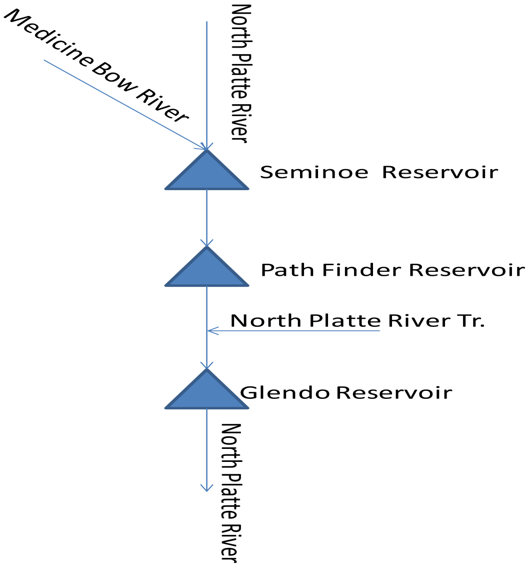

Figure 2. Diagram of North Platte River reservoir system.

3. NORTH PLATTE RIVER RESERVOIRS

Fig. 2 shows the North Platte River, which is approximately 716 miles (1,152 km) long and has a great impact on the agriculture and the general economy of Colorado, Nebraska and Wyoming. The North Platte River is joined by the Medicine Bow River in the Seminoe Reservoir formed by Seminoe Dam. About 50 miles (80 km) downstream the North Platte is joined by the Sweetwater River to form the Pathfinder Reservoir. The North Platte then flows east-southeast across the plains of eastern Wyoming and through the Glendo Reservoir. The Seminoe, Pathfinder and Glendo are the major reservoirs on the North Platte. As a case study, we apply our spatio-temporal data mining method to derive an automated controller for these reservoirs.

3.1 Preprocessing

Each of the reservoirs has a specific set of information associated with it. The water level and rainfall are prevalent in both upstream and the reservoir catchments. These data are recorded using telemetric recorders situated at strategic locations of the reservoirs. The data (in CVS format) was obtained from the U.S Department of Interior Bureau of Reclamation, which recorded at the beginning of each month from January 2000 to May 2011 for each of the reservoirs the following values:

- **Initial storage** (I_t): The amount of water in the reservoir at the beginning of month *t*.
- **End Storage** (S_t): The amount of water in the reservoir at the end of the month *t*.
- **Inflow** (F_t): The amount of water that reaches the reservoir during month *t*.
- **Release** (R_t): The amount of water released from the reservoir during month *t*.

The spatio-temporal data mining algorithm finds functional relationship between the first three variables, which are the input variables, and the last variable, which is the target variable. This relationship can be expressed as follows:

$$R_t = f(F_t, I_t, S_t)$$

All of the upstream reservoirs values influence also the downstream reservoir release values. For example, the release value of the Seminoe reservoir influences the release of the Pathfinder reservoir, while the release of the Glendo reservoir has only a negligible effect on the release of the Pathfinder reservoir. Hence devising operational rules about a given object would require information about the other data objects and their value in the spatial framework.

Data mining has been carried out in JAVA using the WEKA, a commonly used open source machine learning and data mining software. The SMOReg classifier was applied for both the spatial and spatio-temporal dataset. The performance of the quadratic kernel function is compared with the RBF kernel, the Pearson universal kernel and the normalized quadratic kernel. 70% of the data is used for training and the remaining 30% for testing.

Principal component analysis was used to reduce the dimensionality to a much lower number to enhance execution time during the training and testing phase. We have employed MATLAB to generate the PCA components for the spatio-temporal data. Depending on the value of the explained variance, the reduced dimension data is generated.

4. EXPERIMENTAL RESULTS

We examined the effectiveness of the spatio-temporal data mining using SMOReg classifier. In addition, we also studied the capability of different kernel types. We measured the performance efficiency of the classifier using the correlation coefficient and the root mean square error (RMSE) values.

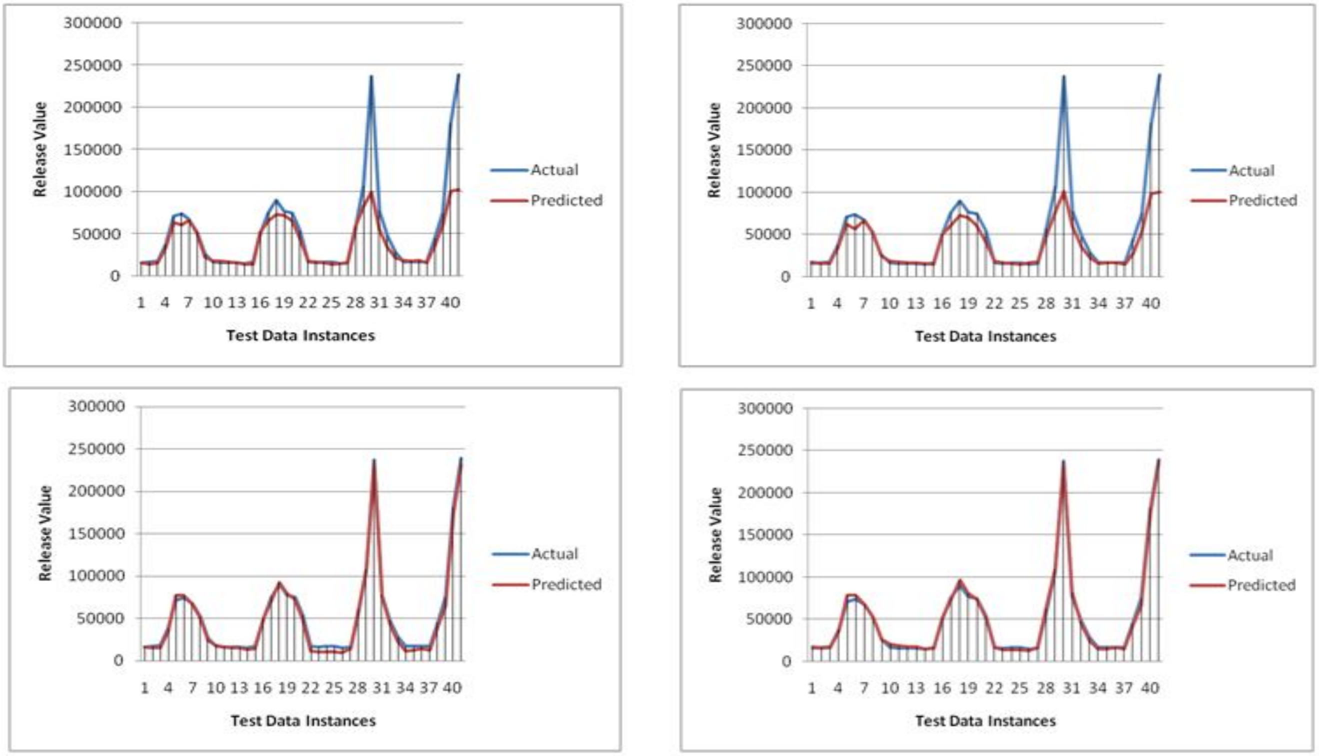

Figure.3. Performance of the Pathfinder Reservoir: quadratic kernel spatial data (top left), quadratic kernel spatio-temporal data (top right), normalized quadratic kernel spatial data (bottom left), and normalized quadratic kernel spatio-temporal data (bottom right).

The efficiency in predicting the release values for the Seminoe, Glendo and Pathfinder reservoirs were recorded. Fig. 3 shows that for the Seminoe reservoir, the best fit was observed with the quadratic and the normalized quadratic kernels. The release values predicted by the model are shown in red and the original recorded release values in blue. The quadratic kernel displays the least prediction errors for the Seminoe reservoir with the reported correlation coefficient value as high as 0.9879 and an RMSE value 3416.91, indicating that it can easily map the input features to result in useful prediction rules. By examining the graphs for the quadratic kernel, it can be stated that in the spatio-temporal model the actual and predicted lines almost overlap. This shows that it is an effective model for the Seminoe reservoir operations. Next reservoir along the river is the Pathfinder Reservoir, for which the quadratic kernel again demonstrates an improvement. The correlation coefficient value was 0.9088 and the RMSE value 6489. The next best fit was observed for the RBF kernel-based data model.

Employing quadratic kernel in the SVR data model for the Glendo Reservoir reports an improvement in performance; increase in correlation coefficient value from 0.9426 to 0.963. However, the best fit is reported for the Glendo reservoi using the RBF kernel with a correlation coefficient of 0.97 and an RMSE value of 10008.9 thus demonstrating improvements with Spatio-Temporal over Spatial Data model

Employing quadratic kernel for the SMOReg classifier reports an improvement in performance, more precisely, an increase in the correlation coefficient value from 0.9426 to 0.963 and decrease in the RMSE value from 17308.9 to 15440.7. The fact that the quadratic kernel gives a better fit for all the Seminoe and the Pathfinder Reservoirs indicates that the true function of the input data vectors can be approximated well by a quadratic function. Therefore, a parametric quadratic model can be concluded as the best fit for predictive spatio-temporal data mining for the Seminoe and the Pathfinder Reservoirs. However, there exists a more complex relationship between the input vectors for the Glendo Reservoir. For the Glendo Reservoir, the model developed with RBF kernel provides the best fit for predicting the release values. That may be due to the fact that the RBF kernel works well for the normalization of the data taking into account the fact that the variability can be reduced by the RBF kernel function, which employs a logarithmic or an exponential transformation of the data such that the resulting data has a normal distribution. Hence, it can be concluded that the quadratic kernel works well when the variance is comparatively high whereas the RBF kernel is more suitable when the variance is relatively small. However, this point needs to be validated with data pertaining to other reservoir systems. Another interesting observation is that there is a high level of compliance between the predicted and the actual values when the release values were typically high and a good compliance between the predicted and the actual when the demand–meeting release values were relatively low. These results show that the support vector regression model performs well.

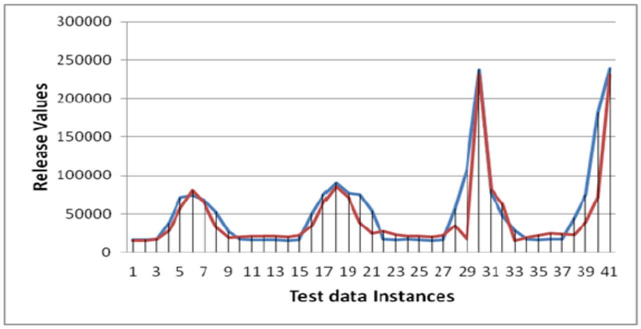

Figure 4. Performance comparison of spatial and spatio-temporal data mining for the Seminoe Reservoir.

4.1 Experiments with Reduced Feature Sets

The aim of PCA is to explain as much of the variance of the observed variables as possible using few composite variables (usually referred to as components), thereby resulting in feature space reduction. For the reservoir dataset, a total of 39 principal components were generated. If 100% of the variance in the correlation matrix were accounted for, then all of the 39 components would need to be retained. However, this would lead to over fitting and would be counter-productive. This prompted the use of explained variance. The performance of the data model for variance between 0.95 and 0.98 is evaluated. The variance values of 0.95 and 0.98 had a set of 4 and 6 principal components, respectively. The data model developed for the Seminoe reservoir with these reduced feature set reported best results with the quadratic kernel based regression model. The RMSE was 22475.2 and the correlation coefficient was 0.93 for variance value of 0.95. Figure 4 shows a comparative evaluation of the predicted and the actual release values. For the Pathfinder Reservoir the same conclusion can be derived, and the best performance was achieved when variance was 0.95. The best fit for the Glendo Reservoir was reported for variance value of 0.98 with correlation coefficient of 0.9456 and RMSE 23497.3.

5. CONCLUSION

Spatio-temporal applications have numerous interactions among the spatial and temporal variables, which must be efficiently modeled for a fully functional data model. In this case study, decision on reservoir water release was investigated for efficient reservoir operation. In this study support vector regression-based data mining technique that employed a quadratic kernel function was applied.

The model performed better on a high-dimensional spatio-temporal data model in comparison with a spatial model. The method was enhanced by reduced-feature support vector regression model with principal component analysis (RF-SVR). The best kernel function and the variance value suitable for a particular reservoir were then carefully determined. For most cases, the quadratic kernel function was able to almost perfectly capture the variable dependency especially in the SVR model except for one case where the RBF kernel reported better results. In the case of the RF-SVR model, the quadratic kernel function proved to be efficient in capturing the relationship to predict the release values. The evaluation results of the RF-SVR data model suggest that dimension reduction with PCA does not drastically decrease the efficiency of SVR-based data models. Though it has some tradeoffs, the RF-SVR proves to be a frontrunner algorithm for modeling spatio-temporal applications.

6. REFERENCES

[1] Anderson, S. and Revesz, P.Z. 2009. Efficient MaxCount and threshold operators of moving objects. *Geoinformatica*, 13, 4 (2009), 355-396.

[2] Cao, H., Mamoulis, N., and Cheung, D. W. 2007. Discovery of periodic patterns in spatiotemporal sequences. *IEEE Trans. Knowledge and Data Eng.* 19, 4 (2007), 453-467.

[3] Chomicki, J. and Revesz, P. Z. 1999. Constraint-based interoperability of spatiotemporal databases. *Geoinformatica,* 3, 3 (1999), 211-243.

[4] Coulibaly, P., Hache, M., Fortin, V., and Bobee, B. 2005. Improving daily reservoir inflow forecasts with model combination. *J. of Hydrological Eng.,* 10, 2 (2005), 91–99.

[5] Dong J, 2000. Research on nonlinear combination forecasting method based on wavelet network. *Journal of Systems Engineering*, 15, 4 (2000), 383-388.

[6] Erwig, M., Güting, R. H., Schneider, M., Vazirgiannis, M. 1999. Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *Geoinformatica,* 3, 3 (1999), 269-296.

[7] Glezakos, T.J., Tsiligiridisa, T.A., Iliadisa, L.S., Yialourisa, C.P., Marisa, F.P., Ferentinosa, K.P. 2009. Feature extraction for time-series data: An artificial neural network evolutionary training model for the management of mountainous watersheds. *Neurocomputing,* 73, 1 (2009), 49–59.

[8] Kanellakis, P. C., Kuper, G. M. and Revesz, P. Z. 1995. Constraint query languages. *Journal of Computer and System Sciences*, 51, 1 (August 1995) 26-52.

[9] Li, L. and Revesz, P. Z. 2004. Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems,* 28, 3 (May 2004), 201-227.

[10] Mennis, J. and Liu, J. W. 2005. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Trans. in GIS*, 9, 1 (2005), 5-17.

[11] Ohashi, O. and Torgo, L. 2012. Wind speed forecasting using spatio-temporal indicators. In *European Conference on Artificial Intelligence*, (August 2102), IOS Press, 975-980.

[12] Revesz, P. Z. 2010. *Introduction to Databases: From Biological to Spatio-Temporal*, Springer, New York, NY.

[13] Revesz, P. Z. 1997. On the semantics of arbitration. *International Journal of Algebra and Computation*, 7, 2 (1997), 133-160.

[14] Revesz, P. Z. and Triplet, T. 2011. Temporal data classification using linear classifiers. *Information Systems*, 36, 1, (2011), 30–41.

[15] Revesz, P. Z. and Triplet, T. 2010. Classification integration and reclassification using constraint databases. *Artificial Intelligence in Medicine*, 49, 2 (2010), 79-91.

[16] Revesz, P. Z. and Wu, S. 2006. Spatiotemporal reasoning about epidemiological data. *Artificial Intelligence in Medicine*, 38, 2 (2006), 157-170.

[17] Solomatine, D . P . and Ostfeld A. 2008. Data-driven modeling: Some past experiences and new approaches. *Journal of Hydroinformatics*, 10, 1 (2008), 3-22.

[18] Tsoukatos, I. and Gunopulos, D. 2001. Efficient mining of spatiotemporal patterns. In *Proc. of the Symp. on Advances in Spatial and Temporal Databases*, Springer, (2001), 425–442.