2009 IEEE International Conference on Data Mining Workshops

A Semi-supervised Framework for Simultaneous Classification and Regression of Zero-Inflated Time Series Data with Application to Precipitation Prediction

Zubin Abraham
Department of Computer Science
Michigan State University
East Lansing, MI-48824, USA
abraha84@cse.msu.edu

Pang-Ning Tan
Department of Computer Science
Michigan State University
East Lansing, MI-48824, USA
ptan@cse.msu.edu

***Abstract*—Time series data with abundant number of zeros are common in many applications, including climate and ecological modeling, disease monitoring, manufacturing defect detection, and traffic accident monitoring. Classical regression models are inappropriate to handle data with such skewed distribution because they tend to underestimate the frequency of zeros and the magnitude of non-zero values in the data. This paper presents a hybrid framework that simultaneously perform classification and regression to accurately predict future values of a zero-inflated time series. A classifier is initially used to determine whether the value at a given time step is zero while a regression model is invoked to estimate its magnitude only if the predicted value has been classified as non-zero. The proposed framework is extended to a semi-supervised learning setting via graph regularization. The effectiveness of the framework is demonstrated via its application to the precipitation prediction problem for climate impact assessment studies.**

***Keywords*-Semi-supervised; Zero-Inflated Time Series Data; Simultaneous Classification and Regression;**

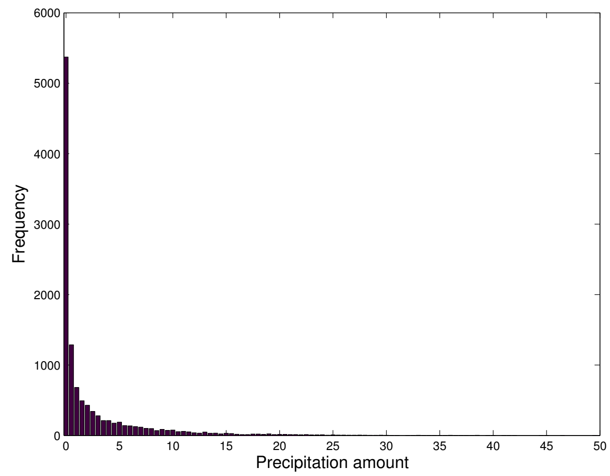

Figure 1. A zero-inflated frequency distribution of daily precipitation at a weather station in Canada

I. Introduction

Predictive models for time series data are commonly employed in the field of economics, finance, epidemiology, ecology, and meteorology, among others. The accuracy of prediction is subject to the choice of model chosen, which in turn may be limited by characteristics of the time series observations and the availability of labeled training data. In the former case, studies have shown that data sets with excess zero values may degrade the performance of classical regression models [2]. Such data are typically encountered in applications such as climate and ecological modeling, disease monitoring, manufacturing defect detection, and traffic accident monitoring. For example, Figure 1 shows the frequency distribution of daily precipitation at a weather station in Canada, in which nearly half of the observations have precipitation values equal to zero. The zero-inflated data often leads to poor fits using standard distributions such as Poisson, exponential, lognormal, and Gamma [16] since they tend to underestimate the frequency of zeros and the magnitude of non-zero values of the data.

While models that cater to zero-inflated data have been developed over the years [2], their accuracies depend on the availability of training data. In recent years, there have been considerable studies directed toward the development of semi-supervised learning methods to utilize unlabeled samples in classification and regression, including for time series prediction problems [6]. However, none of these methods are designed for skewed time series with abundant numbers of zeros.

The objective of this paper is to develop a novel framework to accurately estimate the future values of a zero-inflated time series data by simultaneously performing classification and regression. Unlike other existing zero-inflated models, the framework described by our paper is semi-supervised and not limited by the dimensionality of the data (i.e., number of predictor variables that can be used). Specifically, the proposed framework employs a classification model to determine whether the value at a given time step is non-zero, followed by the application of a regression model to estimate the exact magnitude of the non-zero value. The models are simultaneously learned by optimizing a unified objective function that includes a graph

978-0-7695-3902-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ICDMW.2009.80

regularization term to ensure smoothness of their target functions and consistency between the labeled and unlabeled examples.

In this paper, we demonstrate the effectiveness of our semi-supervised learning framework in the context of precipitation prediction using climate data from the Canadian Climate Change Scenarios Network Web site [1]. We showed that the proposed framework significantly outperforms regression models trained on both zero and non-zero parts of the time series for the majority of the weather stations investigated in this study.

II. Related Work

Time series prediction has long been an active area of research with applications in finance [18], weather forecasting [10], network monitoring [5], transportation planning [13], etc. There are many time series prediction techniques available, including least square regression [15], recurrent neural networks [12], Hidden Markov Model Regression [11], and support vector regression [17]. These techniques are currently employed in a supervised learning setting, and thus, may not fully utilize the value of the unlabeled data.

The motivation behind the use of simultaneous classification and regression framework as opposed to applying regression on the entire time series is due to the zero-inflated data problem. Previous studies have shown that additional precautions must be taken to ensure that the excess zeros do not lead to poor fits [2] of the regression models. A typical approach to model a zero-inflated data set is to use a mixture distribution of the form $P(y|\mathbf{x}) = \alpha\pi_0(\mathbf{x}) + (1-\alpha)\pi(\mathbf{x})$, where π_0 and π are functions of the predictor variables $\mathbf{x}$ and α is a mixing coefficient that governs the probability an observation is a zero or non-zero value. Generally in such an approach, the underlying distribution π may be Poisson or negative binomial distribution (for discrete data) and lognormal or Gamma (for continuous data). These approaches are designed for supervised learning settings and are often limited by the number of explanatory variables that may be used for the prediction task.

There have been extensive studies on the effect of incorporating unlabeled data to supervised classification problems, including those based on generative models[9], transductive SVM [14], co-training [4], self-training [20] and graph-based methods [3]. Blum and Mitchell [4] suggested that unlabeled data can help to reduce variance of the estimator as long as the modeling assumptions match the ground-truth data. Otherwise, unlabeled data may either improve or degrade the classification performance, depending on the complexity of the classifier compared to the training set size [8].

Recently, there have been growing interest on applying semi-supervised learning to regression problems [19]. Cheng and Tan proposed a semi-supervised learning framework for long-term time series forecasting based on Hidden Markov Model Regression [6]. None of these semi-supervised learning methods are designed for handling zero-inflated time series data.

III. Preliminaries

Let $\mathbf{L} = (\mathbf{X}_l, \mathbf{c}'_l)$ be a multivariate time series of length l, where the predictor variables $\mathbf{X}_l = [\mathbf{x}_{l1}, \mathbf{x}_{l2}, ..., \mathbf{x}_{ln}]^T$ is a d-dimensional sequence of values and $\mathbf{c}'_l = [c'_{l1}, c'_{l2}, ..., c'_{ln}]^T$ is the corresponding ground truth values for the response variable. The objective of time series prediction is to learn a target function $f(\mathbf{x}, \mathbf{w})$ that best estimates the future values of the response variable, $\mathbf{c}'_u = [c'_{u1}, c'_{u2}, ..., c'_{um}]^T$, given the historical data $\mathbf{L}$ and the unlabeled data, $\mathbf{X}_u = [\mathbf{x}_{u1}, \mathbf{x}_{u2}, ..., \mathbf{x}_{um}]^T$, where $\mathbf{w} = [w_1, w_2, ..., w_d]^T$ is the set of weights associated with the target function. $\mathbf{X}_u$ may be obtained, for example, using computer-driven simulation models. In the semi-supervised framework proposed in this study, let n represent the number of labeled training points and m the number of unlabeled training points. In the supervised framework proposed, m represents the number of unlabeled testing points.

In this study, we assume the relative frequency of zero values in $\mathbf{c}'_l$ and $\mathbf{c}'_u$ is larger than the frequency of non-zero values. Furthermore, the response variable c' can be mapped into a binary class c, where $c = 1$ if $c' > 0$, and $c = 0$ otherwise. For brevity, we use the notation $y' \equiv f(\mathbf{x}, \mathbf{w})$ as the predicted value of the response variable and y as the predicted class. Let, $\mathbf{y}'_u = [y'_{l1}, y'_{l2}, ..., y'_{lm}]^T$ and $\mathbf{y}_u = [y_{l1}, y_{l2}, ..., y_{lm}]^T$.

In the semi-supervised framework proposed, let $\tilde{\mathbf{y}}$ be a vector of length $n+m$ whose first n elements are initialized with the vector $\mathbf{c}_l$ and whose remaining m elements are initialized with the vector $\mathbf{y}_u$. Hence, in the supervised framework proposed, as there are no unlabeled training points, $\tilde{\mathbf{y}}$ is a vector of length n and is initialized with the vector $\mathbf{c}_l$.

IV. Framework for Simultaneous Classification and Regression

In this paper, we present a framework for predicting future values of a time series with the following unique characteristics:

1) The framework simultaneously performs classification and regression to improve the accuracy of predicting the magnitude of non-zero values in a zero-inflated time series.
2) The framework can be easily extended to a semi-supervised learning setting via graph regularization.

For brevity, we consider only linear regression models, where $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T\mathbf{x}$. Extending the approach to nonlinear models will be a subject for future research.

Our goal is to simultaneously estimate the values of the weight parameters $\mathbf{w}$ and the class labels $\mathbf{y}$ to minimize the following objective function:

$$\arg\min_{\mathbf{w},\mathbf{y}} f(\mathbf{w}) = \sum_{i=1}^{n} c_i (c_i' - y_i y_i')^2 + T_1 \sum_{i=1}^{n} (y_i - c_i)^2 + T_2 \sum_{i,j=1}^{n+m} s_{i,j} [\tilde{y}_i y_i' - \tilde{y}_j y_j']^2 + T_3 ||w||^2$$

where,

$$\sum_d x_{i,d} w_d = y_i'.$$

Intuitively the first term of the objective function is equivalent to the least square formulation of multiple linear regression, except the estimation of $\mathbf{w}$ is performed based on the rain days only. The second term of the objective function measures the classification accuracy on the training data. The third term in the objective function computes the sum of squared difference in the predicted response values for every pair of data points, weighted by the similarity value of their predictor variables. This represents a graph regularization constraint to ensure smoothness of the objective function and can be used to extend the framework to a semi-supervised learning setting. Finally, the last term of the objective function is equivalent to the L_2 norm used in ridge regression models to penalize models that have many large non-zero weights.

Note that each data point corresponds to a given time period in the time series. The similarity matrix S is computed according to the Pearson correlation coefficient between every pair of data points in $\mathbf{X}$. Prior to computing the similarity matrix, each attribute value of the data set is standardized by subtracting the mean value of the attribute and then dividing by its corresponding standard deviation. The standardization of each column is done to account for differences in the variance of the various attributes in the data set. The Pearson correlation value is then transformed to range between 0 and 1. The choice of Pearson correlation as our similarity measure is due to the popularity of the measure in the Earth science domain.

The purpose of the similarity function is to identify how closely related two data points are to one another, and to use this information in creating the regression model which gives more credence to closeness in the predicted amount of precipitation for data points that are similar as against to data points that are dissimilar. As the similarity function has values ranging between 0 to 1, dissimilar data points have limited impact on the error function while similar data points that differ significantly on the amount of predicted precipitation have the largest impact on the error function. The model further emphasizes on using data points that are categorized as rain events by using '0' and '1' as class labels. Such that '0' is assigned to days that are categorized as '`NoRain`' days and '1' to '`Rain`' days.

The supervised version of the framework is obtained by considering only the labeled training examples for the third term in our objective function.

We employ an iterative procedure to solve the objective function. First, we compute the partial derivative of $f(\mathbf{w})$ with respect to each of the w's and set them to zero:

$$\frac{\partial f}{\partial w_k} = \Bigg[2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \Big(\Big(\sum_d \tilde{y}_i w_d x_{i,d} \Big) \Big(\tilde{y}_i x_{i,k} \Big) \Big) + 2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \Big(\Big(\sum_d \tilde{y}_j w_d x_{j,d} \Big) \Big(\tilde{y}_j x_{j,k} \Big) \Big) - 2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \Big(\sum_d \tilde{y}_i \tilde{y}_j w_d (x_{i,d} x_{j,k} + x_{i,k} x_{j,d}) \Big) - 2 \sum_{i=1}^{n} c_i \Big(c_i' - y_i \sum_d w_d x_{i,d} \Big) \Big(x_{i,k} \Big) + 2T_3 w_k \Bigg] = 0$$

This reduces to a system of linear equations of the form $\mathbf{A}\mathbf{x} = \mathbf{b}$ where $\mathbf{x} = [w_1 w_2 .... w_d]^T$ and

$$b_k = \sum_{i=1}^{n} c_i c_i' x_{i,k}$$

A is a square matrix of dimension $d \times d$ where the non-diagonal elements,

$$\mathbf{A}_{k,l} = 2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \tilde{y}_i x_{i,l} x_{i,k} - 2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \tilde{y}_i \tilde{y}_j x_{i,l} x_{j,k} + \sum_{i=1}^{n} c_i y_i x_{i,l} x_{i,k}$$

and the diagonal elements

$$\mathbf{A}_{k,k} = 2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \tilde{y}_i x_{i,k}^2 - 2T_2 \sum_{i,j=1}^{n+m} s_{i,j} \tilde{y}_i \tilde{y}_j x_{i,k} x_{j,k} + \sum_{i=1}^{n} c_i y_i x_{i,k}^2 + T_3$$

Next we apply quadratic discriminant analysis (QDA) on the predicted response values $\mathbf{y}' = \mathbf{w}^T \mathbf{x}$ to estimate the class labels of the unlabeled data points. The updated class labels $\mathbf{y}$ are then used to re-estimate the regression weights $\mathbf{w}$. This procedure is repeated until convergence. A summary of the framework is presented in Algorithm 1. In the remainder of this paper, the supervised and semi-supervised versions of our algorithm are denoted as ZICR-S and ZICR-SS, respectively (where ZICR stands for Zero-Inflated Classification-Regression method).

Algorithm 1 Concurrent Semi-supervised Regression and classification using simultaneous equations iteratively.

Input:
$\mathbf{X}$ (An $(n+m) \times d$ matrix of NCEP weather data)
$\mathbf{c}$ (A n-dimension vector of class labels (1-Rain/0-NoRain))
$\mathbf{c}'$ (A n-dimension vector of precipitation values for each day.)
Output:
$\mathbf{w}$ (A d-dimensional vector of weights)
$\mathbf{y}$ (A $(n+m)$-dimensional vector containing class labels.)
$\mathbf{y}'$ (A $(n+m)$-dimensional vector containing regressional values of amount of precipitation for each day)
Method:
Partition data 3 ways (training, evaluation and test)
1) Perform MLR on the training set (Size-n) to get $\mathbf{w}$.
2) Use the $\mathbf{w}$ on the testing set (Size-m) to get $\mathbf{y}'_i$.
3) Calculate the objective function error using the present $\mathbf{w}$ and save the value
4) Quadratic Discriminant Analysis (QDA) is performed on $\mathbf{y}'_i$ to get $\mathbf{y}_i$
5) In the semi-supervised approach (ZICR-SS) initialize $\tilde{\mathbf{y}}$ to $\mathbf{c}$ for the first n datapoints and initialize the remaining m points of $\tilde{\mathbf{y}}$ with $\mathbf{y}$ from step-3.
In the supervised approach (ZICR-SS), $\tilde{\mathbf{y}}$ is initialized to $\mathbf{c}$ only.
6) Solve $\mathbf{w}$, using the d equations got after differentiating the objective function $f(w)$
7) After having solved $\mathbf{w}$, solve for $\mathbf{y}'$ using the linear equation $\mathbf{y}' = \mathbf{xw}$
8) Apply QDA to find class labels for the training data points $\mathbf{y}$.
10) Calculate the objective function error using the present $\mathbf{w}$
11) For a fixed number of iterations (e.g., 10) or based on the convergence of the objective function, repeat steps 4 to 10
12) Evaluate the model by testing the RMSE error on the test data set.

V. Experimental Evaluation

This section presents the experimental results to demonstrate the effectiveness of our proposed framework.

A. Experimental Setup

The algorithm detailed in the earlier section was applied to climate data from the Canadian Climate Change Scenarios Network Web site [1]. The response variable to be regressed corresponds to daily precipitation values measured at 37 weather stations in Canada. The predictor variables correspond to 26 coarse-scale climate variables derived from the NCEP Reanalysis data set, which include measurements of sea-level pressure, wind direction, vorticity, and humidity, as shown in Table I. The data are available for an entire 40 years period from 1961 to 2001. We also truncate the time series for each weather station to exclude the days in which the precipitation values are missing.

Table I
A Sample of predictor variables for precipitation prediction.

Predictor Variables	
Mean sea level pressure	Surface airflow strength
Surface vorticity	Surface wind direction
Surface divergence	Mean temperature at 2m
500 hPa airflow strength	850 hPa airflow strength
500 hPa zonal velocity	850 hPa zonal velocity
500 hPa meridional velocity	850 hPa meridional velocity

We compared the performance of our algorithm against the multiple linear regression (MLR) model. MLR uses the least square criterion to estimate the weight vector $\mathbf{w}$ of the model. We use the following criteria to evaluate the performance of the models:

B. Experimental Results

1) Performance Comparison: This section compares the RMSE, accuracy, and F-measure values for our proposed supervised (ZICR-S) and semi-supervised (ZICR-SS) framework against the precipitation prediction results of multiple linear regression (MLR). All the experiments were performed using a training size (n) of 3 years starting from the first observation in the time series. The test set size (m) was also fixed at 3 years. After calculating the RMSE on the test set, the training set was shifted by 3 years, such that it now occupied the data set used for testing in the previous iteration. The experiment is repeated 7 times for each station. The RMSE values reported in this section is the mean value of all 7 iterations. The same approach is used to compute the RMSE values for `Rain` days, accuracy (for all days), F-measure for `Rain` days only and F-measure for `NoRain` days only. Due to space restriction, we show the results for 20 weather stations.

As shown in Figure 2, both our models, ZICR-S and ZICR-SS, significantly outperformed the MLR model (trained on all days) in terms of their RMSE values for predicting both `Rain` and `NoRain` days.

The supervised version of our approach outperforming MLR for all 37 stations, while the semi-supervised approach outperformed MLR in 34 out of the 37 stations. In terms of percentage improvement in RMSE, the RMSE for MLR was at an average 8.8% and 8.4% worse than ZICR-S and ZICR-SS respectively. ZICR-S outperformed ZICR-SS in 22 out of the 37 stations.

However, in terms of the RMSE values for `Rain` days only, Figure 3 MLR had an average RMSE value for `Rain` days only that was 4.9% and 5.2% higher than ZICR-S and ZICR-SS respectively. Both ZICR-S and ZICR-SS consistently outperform the MLR model with ZICR-S outperforming in 34 and ZICR-SS outperforming in 32 stations. ZICR-S outperformed ZICR-SS in 21 out of the 37 stations.

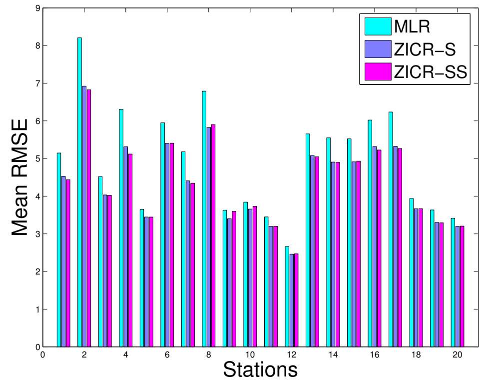

Figure 2. Comparison of RMSE values (for all days) among MLR, ZICR-S, and ZICR-SS.

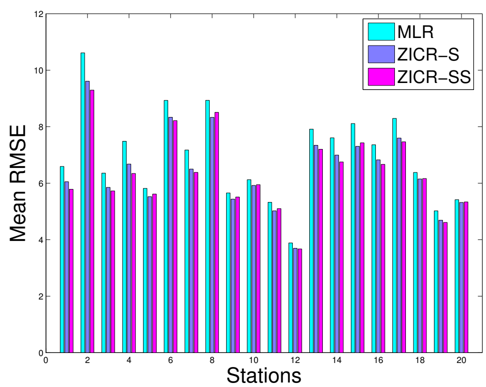

Figure 3. Comparison of RMSE values (for `Rain` days) among MLR, ZICR-S, and ZICR-SS.

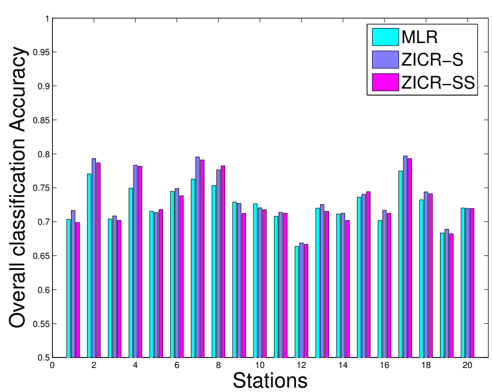

Figure 4. Comparison of classification accuracy (for all days) among MLR, ZICR-S, and ZICR-SS.

Although MLR does not inherently classify any days as `Rain` or `NoRain`, we trained the Quadratic Discriminant Analysis(QDA) classifier used in our framework on the MLR outputs to compare its classification accuracy and F-Measure against those of ZICR-S and ZICR-SS. As shown in Figure 4, all 3 ZICR-S ZICR-SS and MLR were comparable in terms of classification accuracy with ZICR-SS outperforming MLR in approx 60% of the stations. Nevertheless, in terms of F-measure for `Rain` days, both our models consistently outperformed MLR as shown in Figure 5 with ZICR-S outperforming MLR in 32 stations while ZICR-SS outperformed MLR in 33 stations.

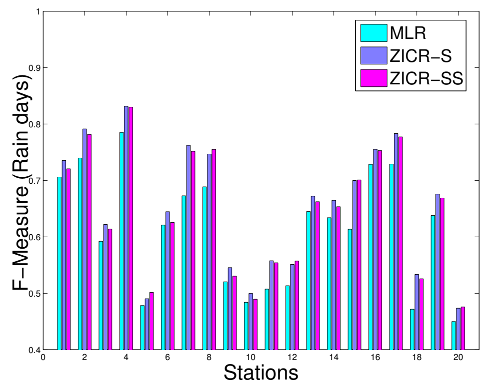

Figure 5. Comparison of F-Measure (for `Rain` days) among MLR, ZICR-S, and ZICR-SS.

With regard to the number of stations that MLR was outperformed in F-measure for `Rain` days, ZICR-S outperformed MLR in 32 and ZICR-SS in 33 stations. Figure 6 shows the comparison of F-Measure for `NoRain` days between MLR, ZICR-S, and ZICR-SS.

VI. Conclusions

This paper presents a novel approach for predicting future values of a time series data that are inherently zero-inflated. The proposed framework decouples the prediction task into two steps—a classification step to predict whether the value of the time series is zero, followed by a regression step to estimate the magnitude of the non-zero time series value. We demonstrate how the framework can be easily extended to a semi-supervised learning setting so as to take advantage of the unlabeled data when there is a dearth of labeled data. The effectiveness of the model was demonstrated on climate data to predict the amount of precipitation at a given station.

The framework presented in this paper assumes a linear relationship between the predictor and response variables. For future work, we plan to extend the framework to capture nonlinear relationships via the use of kernel functions.

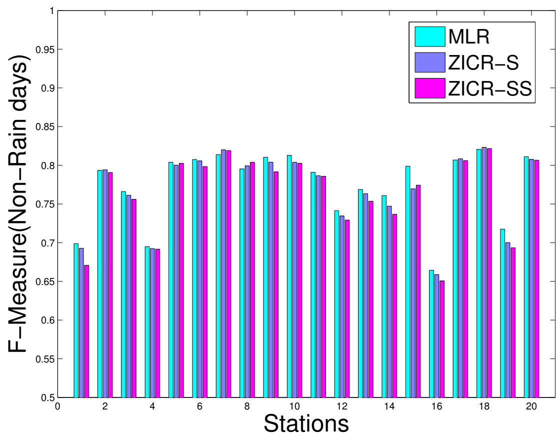

Figure 6. Comparison of F-Measure (for `NoRain` days) among MLR, ZICR-S, and ZICR-SS.

REFERENCES

[1] Canadian Climate Change Scenarios Network, Environment Canada. http://www.ccsn.ca/.

[2] S. Ancelet, M.-P. Etienne, H. Benot, and E. Parent. Modelling spatial zero-inflated continuous data with an exponentially compound poisson process. *Environmental and Ecological Statistics*, DOI:10.1007/s10651-009-0111-6, April 2009.

[3] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proc. of the 18th Int'l Conf. on Machine Learning*, pages 19–26, 2001.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. of the Workshop on Computational Learning Theory*, pages 92–100, 1998.

[5] Y.-A. L. Borgne, S. Santini, and G. Bontempi. Adaptive model selection for time series prediction in wireless sensor networks. *Signal Process*, 87(12):3010–3020, 2007.

[6] H. Cheng and P.-N. Tan. Semi-supervised learning with data calibration for long-term time series forecasting. In *Proc of the ACM SIGKDD Int'l Conf on Data Mining*, Las Vegas, NV, 2008.

[7] I. Cohen, N. Sebe, F. G. Cozman, M. C. Cirelo, and T. S. Huang. Semi-supervised learning of classifiers: Theory and algorithms for bayesian network classifiers and applications to human-computer interaction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(12):1553–1566, Dec 2004.

[8] F. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In *Proc. of the 15th Int'l Florida Artificial Intelligence Society Conference*, pages 327–331, 2002.

[9] F. Cozman, I. Cohen, and M. Cirelo. Semi-supervised learning of mixture models. In *Proc of the 20th Int'l Conf. on Machine Learning*, 2003.

[10] W. Enke and A. Spekat. Downscaling climate model outputs into local and regional weather elements by classification and regression. In *Climate Research 8*, pages 195–207, 1997.

[11] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama. Multiple-regression hidden markov model. In *Proc. of IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, 2001.

[12] C. Giles, S. Lawrence, and A. Tsoi. Noisy time series prediction using a recurrent neural network and grammatical inference. *Machine Learning, 44(1-2)*, pages 161–183, 2001.

[13] W. Hong, P. Pai, S. Yang, and R. Theng. Highway traffic forecasting by support vector regression model with tabu search algorithms. In *Proc. of Int'l Joint Conf. on Neural Networks*, pages 1617–1621, 2006.

[14] T. Joachims. Transductive inference for text classification using support vector machines. In *Proc. of the 16th Int'l Conf. on Machine Learning*, pages 200–209, Bled, SL, 1999.

[15] B. Kedem and K. Fokianos. Regression models for time series analysis. *Wiley-Interscience ISBN: 0-471-36355*, 2002.

[16] T. Martin, B. Wintle, J. Rhodes, P. Kuhnert, S. Field, S. Low-Choy, A. Tyre, and H. Possingham. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters*, 8:1235–1246, 2005.

[17] A. Smola and B. Scholkopf. A tutorial on support vector regression. In *Statistics and Computing*, pages 199–222(24). Spring, 2004.

[18] C.-C. Wong, M.-C. Chan, and C.-C. Lam. Financial time series forecasting by neural network using conjugate gradient learning algorithm and multiple linear regression weight initialization. Technical Report 61, Society for Computational Economics, Jul 2000.

[19] Z. Zhou and M. Li. Semi-supervised regression with co-training. In *Proc. of Int'l Joint Conf. on Artificial Intelligence*, 2005.

[20] X. Zhu. Semi-supervised learning literature survey. In *Technical Report,Computer Sciences, University of Wisconsin-Madison*, 2005.