ISSN : 2231-0711
IJCSET | July 2011 | Vol 1, Issue 6,274-277

An Adaptive Modular Approach to the Mining of Sensor Network Data

Ankit Vyas,Prof.S.K.Satpathy

Department of Computer Science & Engineering
RCET Bhilai Chattishgarh

Abstract: **This paper proposes a two-layer modular architecture to adaptively perform data mining tasks in large sensor networks. The architecture consists in a lower layer which performs data aggregation in a modular fashion and in an upper layer which employs an adaptive local learning technique to extract a prediction model from the aggregated information. The rationale of the approach is that a modular aggregation of sensor data can serve jointly two purposes: first, the organization of sensors in clusters, then reducing the communication effort, second, the dimensionality reduction of the data mining task, then improving the accuracy of the sensing task.**

I. INTRODUCTION

There are plenty of potential applications for intelligent sensor networks: distributed information gathering and processing, monitoring, supervision of hazardous environments, intrusion detection, cooperative sensing, tracking. The ever-increasing use of sensing units asks for the development of specific data mining architectures. What is expected from these architectures is not only accurate modelling of high dimensional streams of data but also a minimization of the communication and computational effort demanded to each single sensor unit. The simplest approach to the analysis of sensor network data makes use of a centralized architecture where a central server maintains a database of readings from all the sensors. The whole analysis effort is localized in the server, whose mission is to extract from the flow of data the high-level information expected to be returned by the monitoring system. If we assume that reasonable-size sensor networks will be made of thousands of nodes, the limitation of this approach is strikingly evident: the number of messages sent in the *Supported by the COMP2SYS project, sponsored by the HRM program of the European Community (MEST-CT-2004- 505079) system as well as the number of variables of the data mining task are too large to be managed efficiently. It has been suggested in literature that alternative architectures are to be preferred in applications where neighboring sensors are likely to have correlated readings [6]. This is the case of aggregating systems which, according to the definition of [10], are systems where the data obtained from the different source nodes can be aggregated before being transmitted along the network. In these systems, we can imagine the existence of intermediary nodes (aggregators) having the capability to fuse the information from different sources. Sensor networks for weather forecasting and monitoring are examples of aggregating systems. The authors of [6] showed that a compression of the sensor information can be performed at local level then reducing the amount of communication and the bandwidth required for the functioning of the system. At the same time techniques of data compression, like Principal Component analysis (PCA) or Independent Component Analysis (ICA) [8], are often used in data mining to reduce the complexity of modeling tasks with a very large number of variables. It is well known in the data mining literature that methods for reducing complexity are beneficial for several reasons: improvement of the accuracy and intelligibility of the model, reduced storage and time requirements. The rationale of the paper is that a modular organization of the sensor network can be used to jointly address the two main issues in mining sensor network data: the minimization of the communication effort and the accurate extraction of high-level information from massive and streaming datasets. In particular this paper proposes a data driven procedure to configure a two-layer topology of a sensor network (Figure 1) made of

1. A lower level whose task is to organize the sensors in clusters, compress their signals and transmit the aggregate information to the upper level,
2. An upper level playing the role of a data mining server which uses the aggregate information to carry out the required sensing task.

We focus here on problems where the sensors are used to perform a supervised learning (e.g. classification, regression or prediction) task: examples could be the classification of traffic fluidity on the basis of route sensing units or the prediction of a wave intensity on the basis of sensors scattered in the ocean. Our approach consists in using a historical data set to find the best way to combine the measures of neighbouring sensors such that the accuracy of the prediction model based on such aggregate measures is optimized. The design procedure relies on an iterative optimization procedure which loops over five steps: (i) a partition of the sensing units in proximity clusters, (ii) the compression of the signals of each cluster of sensors, (iii) the aggregation and transmission of the compressed signals to the upper data mining server, (iv) the training of the prediction model in the data mining server, and (v) the assessment of the partition according to multiple criteria, like the prediction accuracy of the data mining model and the energy and transmission requirements of the resulting network. The best partition which is returned by this multi-criteria optimization procedure can be used as a template for the topological organization of sensors. An important issue in mining sensor network data concerns the streaming and possibly non stationary nature of data. Non stationary may be

due to changes in the phenomenon underlying the measures as well to sensor malfunctioning and/or modifications of their geographical location. In order to address this aspect we have recourse to adaptive features at both levels of our architecture. At the lower sensor integration level we use an effective sequential implementation of the Principal Component Analysis (PCA) technique: the PAST algorithm [11]. The upper data mining module uses an adaptive lazy learning (LL) technique [1] characterized by a fast training phase and an effective treatment of non stationary. The experimental section of the paper presents some preliminary results obtained by adopting the proposed two-layer architecture in the context of a simulated monitoring task: measuring the wavelength of a two dimensional wave in situation of scattering.

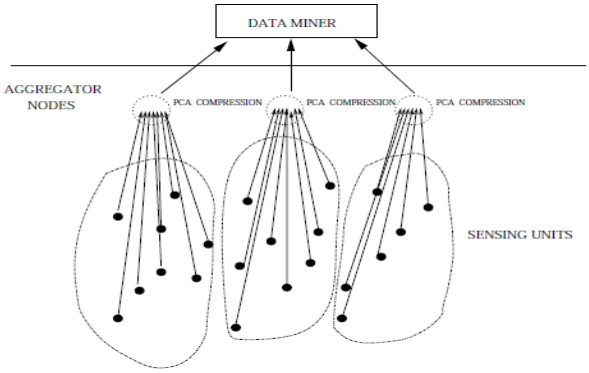


Fig 1: The black dots represent the sensing units. The dotted circles represent the aggregator nodes which carry out the fusion of data coming from neighbouring sensors before sending the aggregated signals up to the data mining server

II. PROBLEM IDENTIFICATION

Consider a sensor network S made of S sensors where P is a [S, 3] matrix containing the three-dimensional coordinates of the S sensors and (2.1) x(t) = {s1(t), s2(t), . . . , sS(t)} is the state (or snapshot) of the sensor network at time t. Suppose we intend to employ S to perform a supervised learning task, for example a regression problem

(2.2) y(t) = f(x(t)) + ε(t) where y is the variable to be predicted at time t on the basis of the state x(t) of the network S and ε is usually thought as the term including modeling error, disturbances and noise. If we have available a finite dataset DN = {_x(ti), y(ti)_, i = 1, . . . , N} of N input-output observations, this problem can be tackled as a conventional regression problem, by first estimating an approximator of f on the basis of DN and then using this estimator as a predictor of y. However, if, like in the case of sensor networks, the number S is huge, the mapping f is non-stationary and the data are collected sequentially, conventional techniques reach rapidly their limits. In particular, the large dimensionality of the problem asks for feature selection problem as well as the streaming aspect of the problem requires sequential (also called recursive) estimation approaches. This paper proposes an approach to the problem of data mining in sensor networks which tries to conciliate the needs for an accurate prediction of the output y with the constraints related to energy reserves, communication bandwidth and sensor computational power. The following subsections will rapidly sketch the two computational modules used in our approach: the recursive PCA and the adaptive Lazy Learning. Section 5 will describe how these modules are combined in our architecture for mining sensor networks.

III. PCA COMPRESSION TECHNIQUES

As discussed above, each data mining problem in the context of sensor network data with large S has to face the problem of reducing dimensionality. Existing techniques for feature selection (for an up-to- ate state of the art on feature selection see [7]) can be grouped into two main approaches: the wrapper and the filter approach. In the wrapper approach [9] the feature subset selection algorithm exists as a wrapper around the learning algorithm, which is often considered as a black box able to return (e.g. via cross-validation) an evaluation of the quality of a feature subset. On the contrary, the filter approach selects features using a pre-processing step independently of the learning algorithm. In this paper we will adopt the Principal Component analysis (PCA) technique, an instance of the filter approach. PCA is a classic technique in statistical data analysis, feature extraction and data compression [8]. Given a set of multivariate measurements, its goal is to find a smaller set of variables with less redundancy, that would give as good a representation as possible. In PCA the redundancy is measured by computing linear correlations between variables. PCA entails transforming the n original variables x1, . . . , xn into m new variables z1, . . . , zm (called principal components) such that the new variables are uncorrelated with each other and account for decreasing portions of the variance of the original variables. Consider a vector x of size n and a matrix X containing N measures of the vector x. The m principal components

$$(3.3) \quad z_k = \sum_{j=1}^{n} w_{jk} x_j = w_k^T x, \qquad k = 1, \ldots, m$$

are defined as weighted sums of the elements of x with maximal variance, under the constraints that the weights are normalized and the principal components are uncorrelated with each other. It is well-known from basic linear algebra that the solution to the PCA problem is given in terms of the unit-length eigenvectors e1. . . en of the correlation matrix of x. Once ordered the eigenvectors such that the corresponding eigenvalues λ1, . . . , λn satisfy λ1 ≥ λ2 ≥ . . . ≥ λn, the principal component zk is given by zk = eT k x. It can be shown that the PCA problem can be also put in the form of a minimum mean-square error compression of x. This means that the computation of the wk for the first m principal components is equivalent to find the orthonormal basis w1, . . . , wm that minimizes

$$(3.4) \quad J_{PCA} = \frac{1}{N} \sum_{t=1}^{N} \|x(t) - \sum_{k=1}^{m} (w_k^T x(t)) w_k\|^2$$

If we denote W = (w1, . . . , wm)T where W is a matrix of size [m, n] we have

$$(3.5) \qquad J_{PCA} = \frac{1}{N} \sum_{t=1}^{N} \|x(t) - W^T W x(t)\|^2$$

What is appealing in this formulation is that a recursive formulation of this least-squares problem is provided by the Projection Approximation Subspace Tracking (PAST) algorithm proposed by [11]. This algorithm, based on the recursive formulation of the least squares problem, has low computational cost and makes possible an updating of the principal components as new observations become available. Once the matrix W is computed a reduction of the input dimensionality is obtained by transforming the input matrix X into the matrix Z = XWT and by transforming the regression problem of dimensionality n into a problem of dimensionality m in the space of principal components. At this step the question arises of how to choose m. The techniques more commonly used rely either on the absolute values of the eigenvalues or on procedures of cross-validation [8].

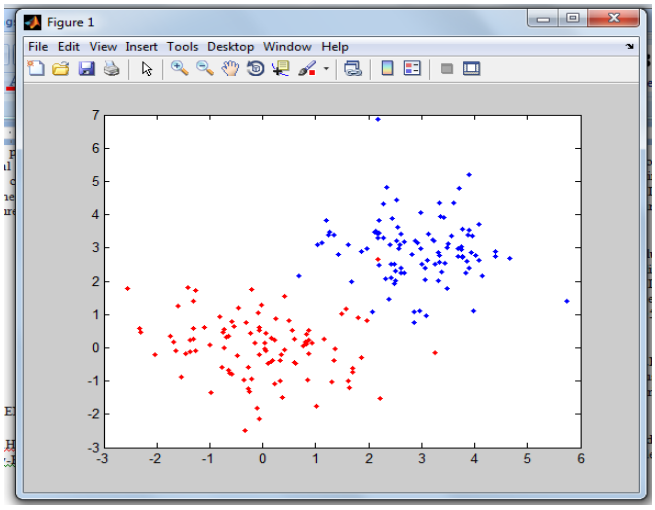


Fig2: plot training examples

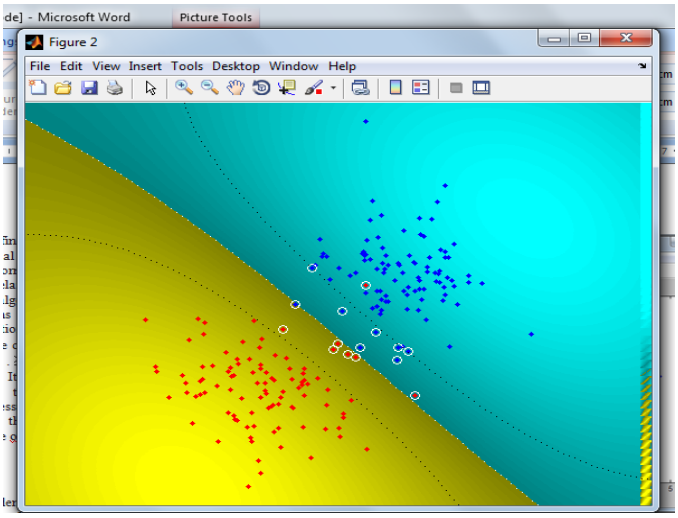


Fig3: Training SVC

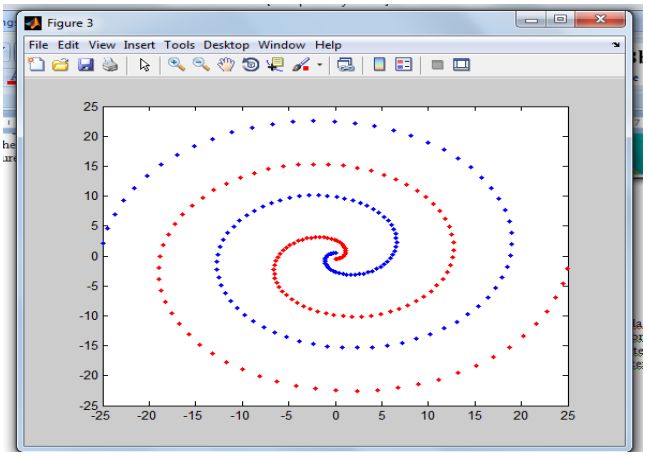
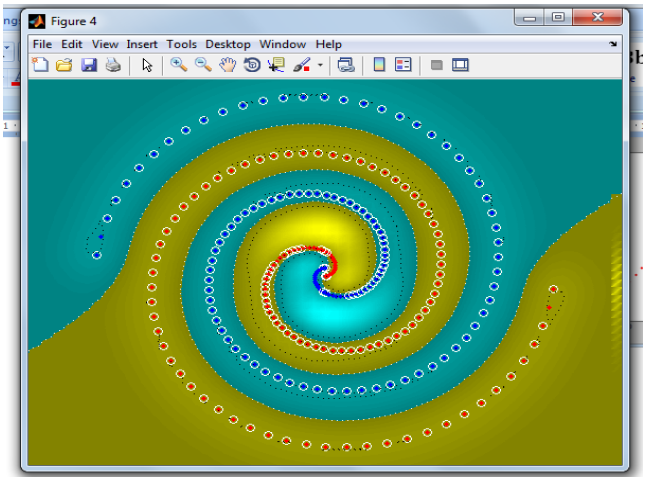


Fig 4: plot training examples after Training SVC



Fihg 5 :Ploting Training results,onvalid for 2-dimensional examples

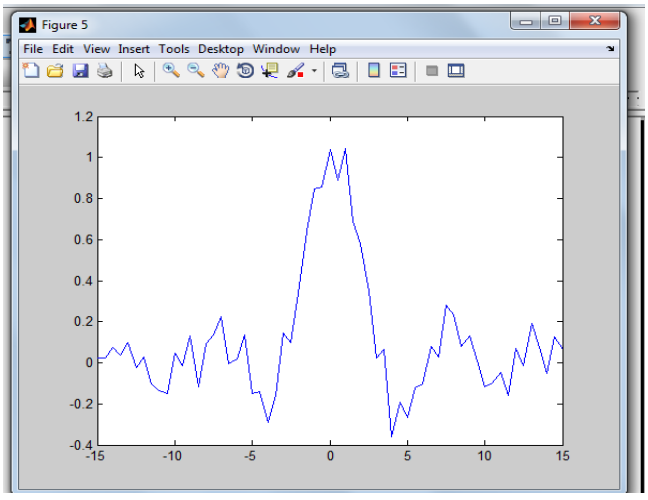


Fig 6 :support vector machines for regression

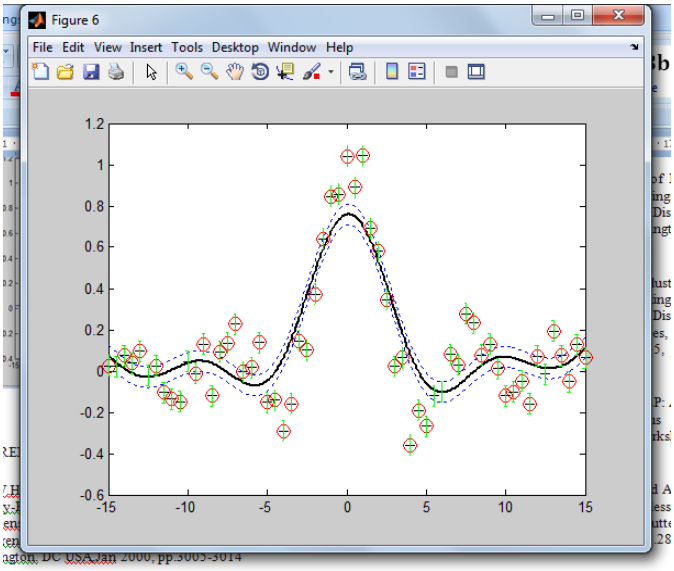


Fig 7: Pridict the output for test examples Function regration

CONCLUSIONS

In this paper, we have presented a framework for building and deploying predictors in sensor networks in a distributed way, by building local models at the sensors and transmitting target class predictions rather than raw data to the root. This framework is appropriate for the limited resources found in sensor networks, due to power, bandwidth and computational limits. We have also showed how the use of local predictive models enables sensors to respond to changes in targets by relearning local models when their local predictive accuracy drops below a threshold. This enables efective distributed data mining in the presence of moving class boundaries, and also creates new possibilities, for example the use of sensors to detect anomalies, even when the criteria for an anomaly changes over time. Finally, because only model predictions rather than data are transmitted, the framework is also suitable for settings where data confidentiality is a concern.

REFERENCES

[1] W.Heinzelman,A.Chandrakasan, H. Balakrishnan. "Energy-Efficient communication protocol for wireless microsensor network", Proc. of the Hawaii International Conference on System Sciences, IEEE Computer Society, Washington. DC USA,Jan 2000, pp.3005-3014

[2] S. Lindsey, C. S. Raghavendra, "PEGASIS: Power-Efficient gathering in sensor information systems". In: Proc. of the IEEE Aerospace Conference, Big, Sky, Montana, July 2002, vol.3, pp.1125-1130

[3] A. Manjeshwar and D. P. Agarwal, "TEEN: a routing protocol for enhanced efficiency in wireless sensor networks", In proceedings of the 1st International Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, IEEE Computer Society, San Francisco, April 2001, pp.2009-2015

[4] M. J. Handy, M. Haase and D. Timmermann, "Low energy adaptive clustering hierarchy with deterministic cluster-head selection", IEEE International Conference on Mobile and Wireless Communications Networks, IEEE Communications Society, Stockholm, 2002, pp.368-372

[5] C.M Liu, C. H. Lee, L. C. Wang, "Distributed clustering algorithms for data-gathering in wireless mobile sensor networks", Parallel and Distributed Computing, Academic Press, Inc. Orlando FL USA, 2007, pp.1187-1200

[6] C. F. Li, M. Ye, G.h. Chen and J. Wu, "An energy-efficient unequal clustering mechanism for wireless sensor networks", In Proc. of 2nd IEEE International Conference on Mobile Ad-hoc and Sensor Systems, Washington DC, November 7-10 2005, pp.597-604

[7] Y.Wang, M.Xu, "Monte Carlo simulation of LEACH protocol for wireless sensor networks", Proceedings of the Sixth International Conference on Parallel and Distributed Computing, IEEE Computer Society, Washington DC USA,2005, pp.85-88

[8] L. Ying, H. B. Yu, "Energy adaptive cluster-head selection for wireless sensor networks", Proceedings of the Sixth International Conference on Parallel and Distributed Computing Applications and Technologies, IEEE Computer Society, Washington DC USA, 2005, pp.634-638

[9] G. Smaragdakis, I. Matta, A. Bestavros, "SEP: A stable election protocol for clustered heterogeneous wireless sensor networks", In: Proc. of the Int'l Workshop on SANPA, Boston Massachusetts USA, 2004

[10] D. J. Dechene, A. E. Jardali, M. Luccini and A. Sauer, "A Survey of Clustering Algorithms for Wireless Sensor Networks", Computer Communications, Butterworth-Heinemann Newton, MA USA, October 2007,pp.2826-