Procedia Computer Science

Volume 29, 2014, Pages 2201–2207

ICCS 2014. 14th International Conference on Computational Science

Autonomous Framework for Sensor Network Quality Annotation: Maximum Probability Clustering Approach

Ritaban Dutta[1*], Aruneema Das[2], Daniel Smith[1], Jagannath Aryal[2], Ahsan Morshed[1], Andrew Terhorst[1]

[1] *CSIRO Computational Informatics, Hobart, Tasmania, Australia*
2 University of Tasmania, Hobart, Tasmania, Australia
Ritaban.Dutta@csiro.au

Abstract

In this paper an autonomous feature clustering framework has been proposed for performance and reliability evaluation of an environmental sensor network. Environmental time series were statistically preprocessed to extract multiple semantic features. A novel hybrid clustering framework was designed based on Principal Component Analysis (PCA), Guided Self-Organizing Map (G-SOM), and Fuzzy-C-Means (FCM) to cluster the historical multi-feature space into probabilistic state classes. Finally a dynamic performance annotation mechanism was developed based on Maximum (Bayesian) Probability Rule (MPR) to quantify the performance of an individual sensor node and network. Based on the results from this framework, a "data quality knowledge map" was visualized to demonstrate the effectiveness of this framework.

Keywords: Maximum (Bayesian) Probability Rule (MPR), PCA, FCM, SOM, Sensor Network.

1 Introduction

The data availability from a sensor network is often very limited and data quality is subsequently very poor. This practical limitation could be due to the difficult geographical location of a sensor node, or sensor station, extreme environmental conditions, communication network failure or technical failure of the sensor node. Data uncertainty from a sensor network makes the network unreliable and inefficient. This inefficiency leads to the failure of natural resource management systems, such as agricultural water resource management, weather forecast, crop management, or irrigation scheduling. Inefficient performance evaluation and poor data quality often result in expensive maintenance, and eventual high risk of network decommissioning. This study focuses on

[*] Masterminded the work and corresponding author of this paper

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2014

doi:10.1016/j.procs.2014.05.205

developing an autonomous sensor network quality annotation framework to tackle this problem in an inexpensive software based approach. The intended purpose of this proposed Autonomous Maximum Probability Clustering (AMPC) Framework is to provide an automatic timely quality trend detection system for a real life sensor network (Dutta 2013, Nasipuri 2006, Bartsch 1996).

2 South Esk Sensor Web

The Sensor Web is an advanced spatial data infrastructure in which different sensor assets can be combined to create a macro-instrument of sensing capability. This macro-instrument can be instantiated in many ways to achieve multi-modal observations across different spatial and temporal scales.

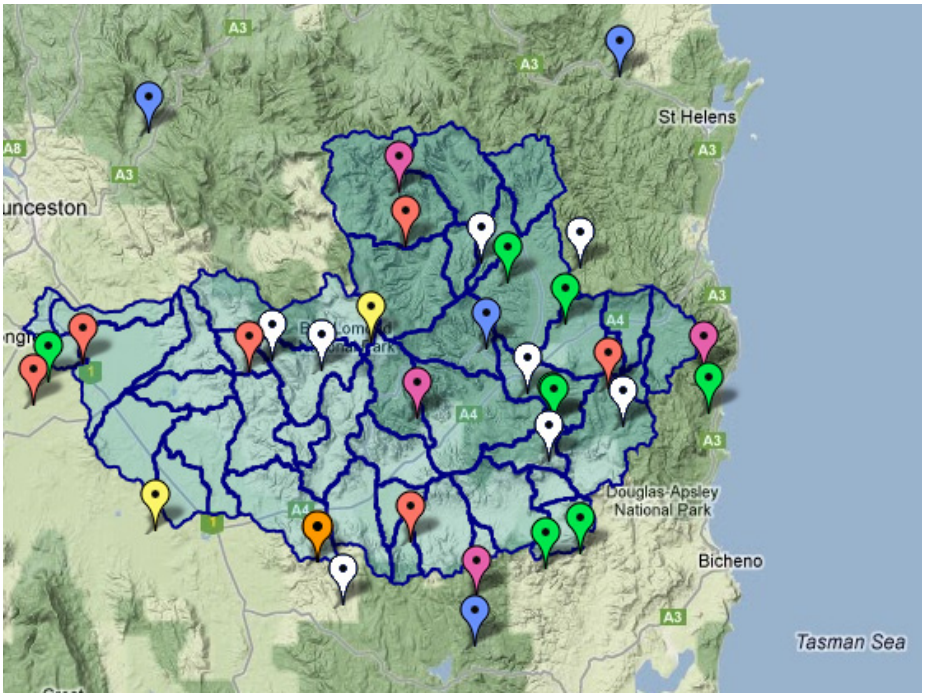

Figure 1: The Google Maps™ pane below presents a federated view of near real-time sensor data from the different sensor networks operating in the South Esk catchment (South Esk Website 2013).

CSIRO is investigating how emerging standards and specifications for Sensor Web Enablement can be applied to the hydrological domain. To this end, CSIRO is implementing a Hydrological Sensor Web in the South Esk river catchment in NE Tasmania (Figure 1). The South Esk river catchment was chosen because of its size (3350km^2, large enough to show up differences in catchment response to rainfall events), spatial variability in climate (there is an 800mm range in average annual rainfall across the catchment), fickle nature of seasonal flow, and relatively high-level of instrumentation.

This is made possible by re-publishing near real-time sensor data provided by the Bureau of Meteorology (BoM), Hydro Tasmania, Tasmania Department of Primary Industries, Parks, Wildlife and Environment (DPIPWE), Forestry Tasmania and CSIRO via a standard web service interface (Sensor Observation Service) developed by the Open Geospatial Consortium (OGC). The Sensor

Observation Service implementation that CSIRO is using was developed by 52North. Exposing sensor data via standard web service interfaces provides a much richer picture of what is going on in the catchment (enhanced situation awareness).

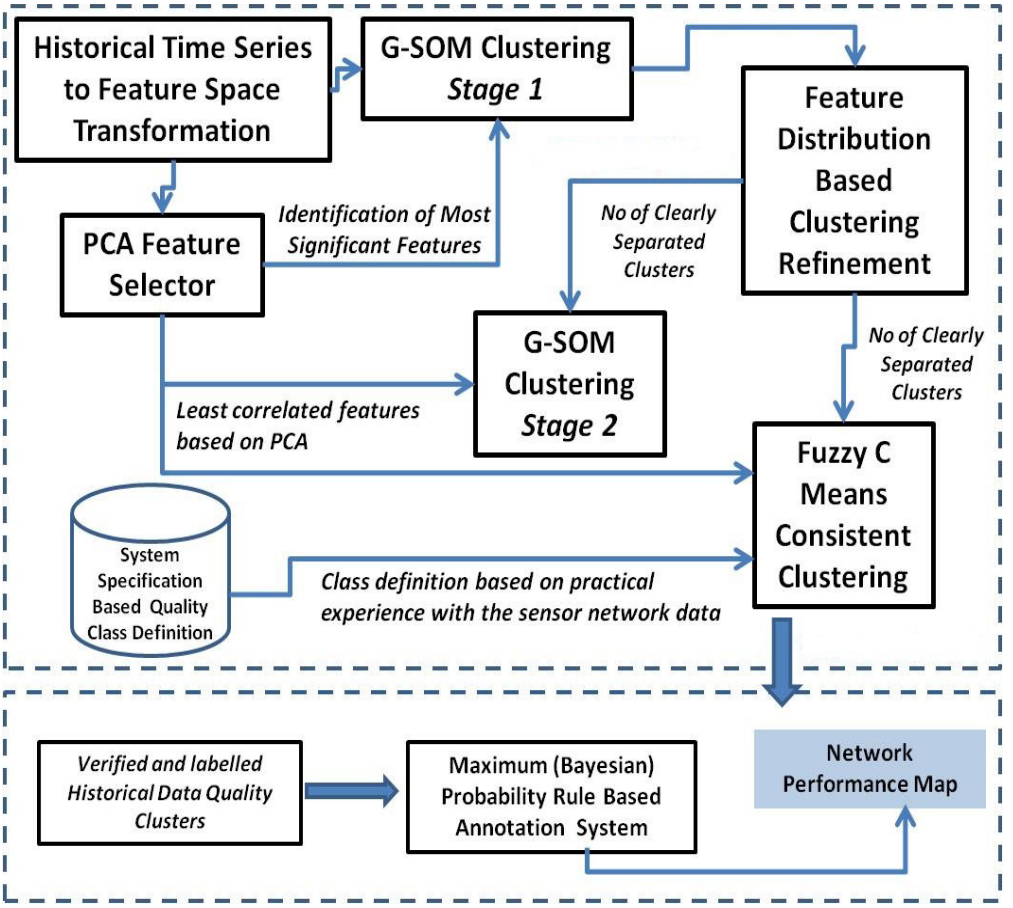

Figure 2: Workflow diagram of the AMPC framework.

3 Time Series Feature Extraction

Data from the sensor network node was with half hourly resolution. The AMPC framework was developed using historic two year's air temperature data (^{o}C) recorded from a single sensor network node. On any given day, a moving window of 600 days was used to developed off-line static autonomous knowledge discovery model, which was then used in the dynamic phase to annotate the quality of newly available data from the same sensor node. This fixed window provided fixed computational cost for the framework. Multiple statistical features were extracted from the individual time series recorded from a sensor node. This approach helped to reduce the dimensionality of observation points for analysis purposes and to optimize the required computational power. A daily time window was used for segmentation (48 data points in one 24 hrs segment) purposes before multiple feature extraction. The daily extracted feature set included, a) maximum value, b) minimum value, c) number of days with missing value, d) standard deviation, e) length of largest consecutive missing value part, f) Kurtosis, g) Skewness, h) area under curve and i) number of peaks two hours

apart. The skewness of a distribution is defined as, $y = \frac{E(X-\mu)^3}{\sigma^3}$ and the kurtosis of a distribution is defined as, $y = \frac{E(X-\mu)^4}{\sigma^4}$ where μ is the mean of x, and σ is the standard deviation of x, and E(t) represents the expected value of the quantity t. The number of peaks was estimated during the whole day. The area under the segmented curve was calculated using simple data integration, with the smallest step as 30 min (Dutta 2013, Das 2009, Buonadonna 2005).

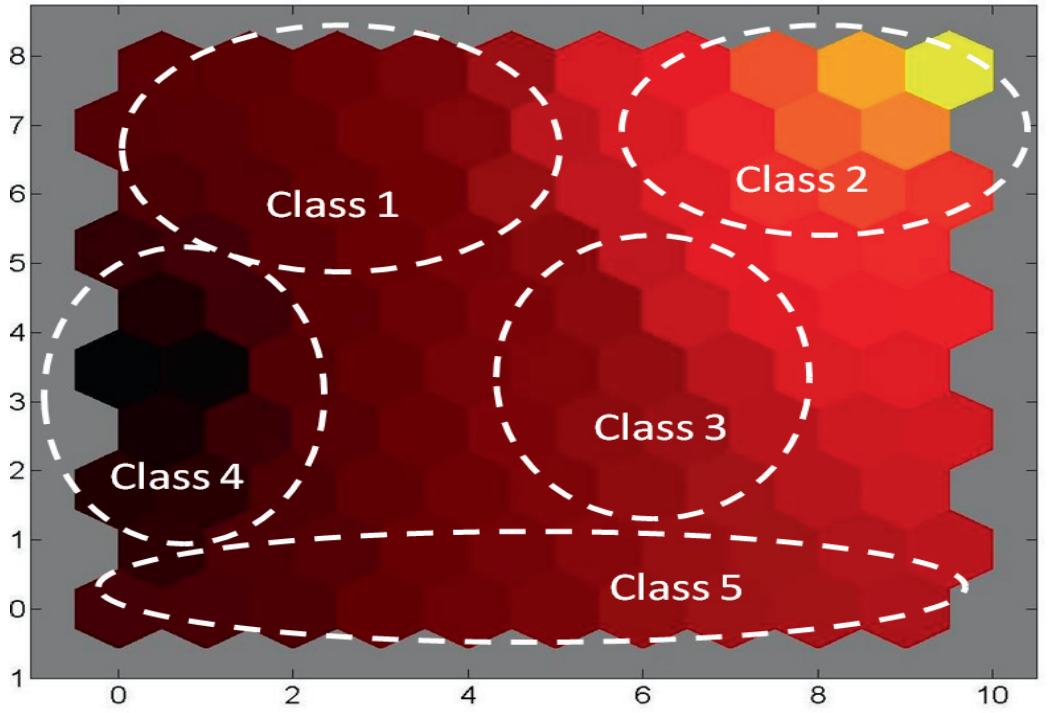

Figure 3: SOM clustering guided by least correlated features. Positions of the most significant clusters were estimated automatically based on the distributions of feature points associated with the trained nodes.

4 Clustering Framework and Analytics

Figure 2 shows the framework workflow diagram of the AMPC framework. The autonomous clustering based knowledge discovery AMPC consists of a PCA, G-SOM and an FCM processing blocks. These functional blocks discover, cross validate and express meaningful natural group of clusters about the data quality trends over long period of time, whereas the dynamic annotation function has the MPR based classification system, which annotates new data.

The PCA is an unsupervised linear non-parametric projection method used for dimension reduction and feature extraction. PCA was applied to extract least correlated features and the maximum data variances. PCA feature selector found that the 'area under curve', 'maximum value', and 'number of days with missing value' features were the most significant; representing 95% data variance along first three components. In Stage 1 selected PCs were used to guide G-SOM clustering with a pre-defined large grid of 100 neurons in conjunction with an iterative feature distribution based cluster refinement

optimizer. Figure 3 shows the Stage 1 G-SOM clustering and refinement outcomes with minimum inter cluster overlapping (Yuxi 2009, Horling 2009, Das 2009).

Weights of trained grid nodes were used to determine the optimum number of clusters and their best positions. G-SOM clustering identified five natural clear clusters (C = 5) with minimum overlapping, where 90% of the feature points were concentrated, by calculating total feature points associated with each neuron. Autonomously detected cluster number was used to initialize the Stage 2 G-SOM training grid of size 5 X 1.

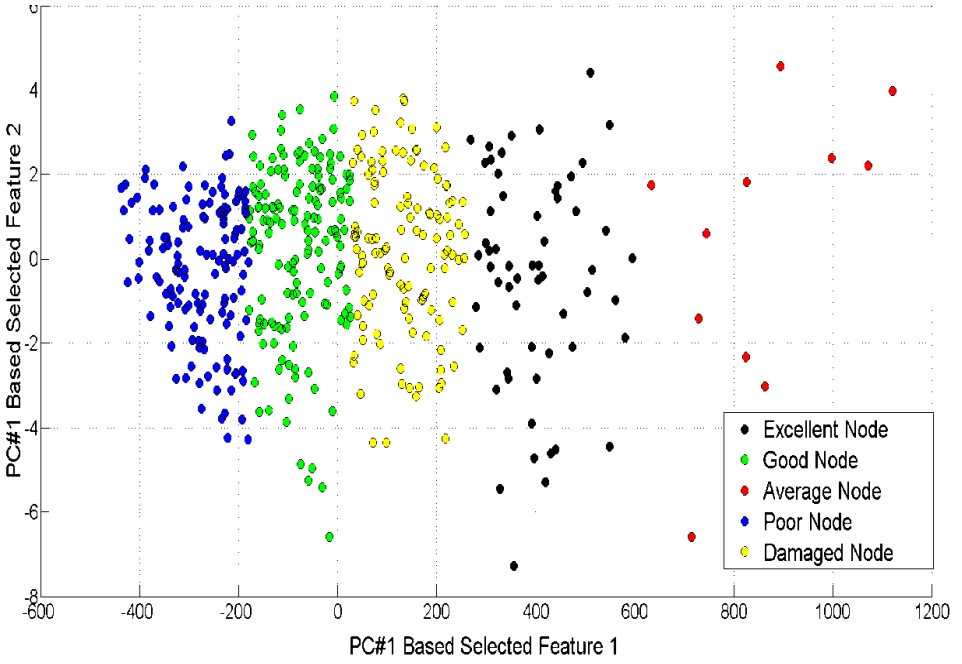

Figure 4: PCA-SOM-mFCM based historical "data quality knowledge map" system. This map system was used in the MPR based classification and dynamic quality annotation purposes.

Independent validation and labeling of these clusters was performed using m-FCM clustering. Ten parallel FCM blocks were applied to find most consistent clusters as they were guided by the Stage 1 G-SOM cluster refinement. The most consistent 5 clusters independently optimized and identified based on m-FCM were then labeled using the pre-defined system specification based sensor node data quality classes; namely {'Excellent Node', 'Good Node', 'Average Node', 'Poor Node', 'Damaged Node'}. Figure 4 shows probabilistic "data quality knowledge map" determined by the AMPC framework. This node based knowledge clustering system was used in the MPR based quality classification and annotation system (Hand 1981, Dutta 2004). The posterior and prior probabilities of class membership were correlated using typicality probability of class by Baye's Rule (Dutta 2006). An Adaptive Kernel Estimator was applied, which formed the MPR as the following expression:

$$\hat{P}(g / \mathrm{X_u}) = \frac{q_g . \hat{f}(X_u / g)}{\sum_{g=1}^{k} q_{g'} . \hat{f}(X_u / g')}$$

The probability denoted by P(g/Xu) is the probability of unit u belonging to group g, given that the unit has a particular observation vector X_u, is called the "posterior probability". The decision rule about the dynamic data quality annotation becomes {Assign unit u to class g if $q_g.\hat{f}(X_u / g) > q_h.\hat{f}(X_u / h)$, for g $\neq$ h}, where $q_g = \hat{\pi}_g$, and h is global smoothing factor for the quality classes. The AMPC framework was applied to dynamically classify any new sensor observation into one of these quality clusters. Splitting the framework into static knowledge modelling and dynamic annotation phases made this approach computationally efficient and truly autonomous. The 'area under curve', 'maximum value', and 'number of days with missing value' features were used for Figure 4. Framework was able to classify unknown (independently labeled) data with 85% accuracy, 90% specificity and 87% sensitivity.

5 Conclusion

On the basis of data availability, data preprocessing and interpolation results a dynamic time series annotation system was developed to provide recommendation about the South Esk sensor web data. Individual time series was labeled as {'Excellent Node', 'Good Node', 'Average Node', 'Poor Node', and 'Damaged Node'} categories. Processed time series were stored in a data structure along with recommendation. The processed data were also encapsulated with extracted statistical features and extracted meta information i.e. maximum value event and the date of that even, minimum value event and the date of that event, largest missing value segment with corresponding dates, maximum number of consecutive days with least data variance. All this processed information were part of dynamic data annotation system. Idea was to process the time series data dynamically, annotate, and provide a generic data usability recommendation about the data for any environmental application.

An autonomous approach of this newly designed AMPC framework could form a benchmarking solution for any real time sensor network's data quality annotation system.

References

Dutta R, Morshed A, Performance Evaluation of South Esk Hydrological Sensor Web: Unsupervised Machine Learning and Semantic Linked Data Approach, IEEE Sensors Journal, (2013), Vol 13, Issue: 10, 3806 - 3815.

Dutta R, Smith D, Timms G, Dynamic Annotation and Visualisation of the South Esk Hydrological Sensor Web, IEEE ISSNIP, Melbourne, Australia, pp 105-110, 2013.

Dutta R, Dutta R, "Maximum probability rule" based classification of MRSA infections in hospital environment: Using electronic nose, Sensors and Actuators B: Chemical (2006), 120, pp 156-165.

Nasipuri A, and Subramanian K, Development of a Wireless Sensor Network for Monitoring a Bioreactor Landfill, GeoCongress 2006. http://www.ece.uncc.edu/~anasipur/ pubs/geo06.pdf

Bartsch M, Weiland T., and Witting M, Generation of 3D isosurfaces by means of the marching cube algorithm, Magnetics, IEEE Transactions on Volume 32, Issue 3, Part 1, pp.1469 – 1472, May 1996

The South Esk website. [Online] (2014).

Available: http://www.csiro.au/sensorweb/au.csiro.OgcThinClient/OgcThinClient.html

Das A, Stocks N G, Hines E L, Enhanced coding for exponentially distributed signals using suprathreshold stochastic resonance, Elsevier Communications in Nonlinear Science and Numerical Simulation, 14, pp 223-232, 2009.

Buonadonna P, Gay D., Hellerstein J M, Hong W, and Madden S, TASK: sensor network in a box, Proceedings of the Second European Workshop on Wireless Sensor Networks, pp 133 -144, 2005.

Yuxi H, Deshi L, Xueqin H, Tao S, Yanyan H, The Implementation of Wireless Sensor Network Visualization Platform Based on Wetland Monitoring, Second International Conference on Intelligent Networks and Intelligent Systems, pp 224- 227, Nov 2009.

Horling B, Vincent R, Mailler R, Shen J, Becker R, Rawlins K, Lesser V, Distributed sensor network for real time tracking, Second International Conference on Intelligent Networks and Intelligent Systems, pp224- 227, Nov 2009.

Hand D.J., Discrimination and classification, (1981) New York: Wiley.

Dutta R, Gardner J W Gardner, Hines E L, Classification of ear, nose, and throat bacteria using a neural-network-based electronic nose, MRS bulletin 29 (10), pp 709-713, 2004.

Das A, Stocks NG, Hines EL, Enhanced coding for exponentially distributed signals using suprathreshold stochastic resonance, Elsevier Communications in Nonlinear Science and Numerical Simulation, 14, pp 223-232, 2009.