Scientific Research and Essays Vol. 6(27), pp. 5685-5695, 16 November, 2011
DOI: 10.5897/SRE11.333
Available online at http://www.academicjournals.org/SRE
ISSN 1992-2248 ©2011 Academic Journals

Full Length Research Paper

Anomaly detection and prediction of sensors faults in a refinery using data mining techniques and fuzzy logic

Mahmoud Reza Saybani*, Teh Ying Wah, Amineh Amini and Saeed Reza Aghabozorgi Sahaf Yazdi

Department of Information Science, Faculty of Computer Science and Information Technology, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia.

Accepted 18 October, 2011

Like all manufacturing companies, refineries use many sensors to monitor and control the process of refining, therefore it is very crucial to detect any sensor faults or anomalies as early as possible, and to be able to replace or repair a sensor well in advance of any fault. Objective of this paper is to present a method for detecting anomalies in a sensor data, as well as to predict next occurance of a sensor failure. Data mining techniques to detect anomaly in sensor data and predict the occurrence of next faulty event were introduced. For anomaly detection, this research used MATLAB's fuzzy logic toolbox tools to find clusters which uses subtractive fuzzy clustering algorithm and generates a model, a Sugeno-type fuzzy inference system. The same toolbox was used to evaluate the model with a promising result. To predict sensor fault, the original time series were used to create a new 'derived time series'. Two prediction models known as 'auto regressive integrated moving average' and 'autoregressive tree models' were used against the new time series to predict next occurrence of sensor failure. The results of these models were compared. The model developed and introduced in this paper serves as an additional tool, which helps not only engineers and operators of oil refineries, but also other engineers of other disciplines to work more efficiently.

Key words: Data mining, derived time series, sensor fault detection and prediction, fuzzy clustring, machine learning, oil refinery.

INTRODUCTION

With thousands of sensors in use in a refinery, it is very time consuming and labor intensive to keep track of whether they work properly or not, even through a cyclic maintenance of sensors, it may happen that the faults are not detected during the maintenance. Therefore it is crucial to determine a proper time for maintenance of sensors or systems. Due to increasing use of computers and cheaper storage media, and faster computing, manufacturing companies such as refineries have gathered a lot of data. Dealing with the huge amount of data is out of reach for the operators, specially when it comes to detecting useful patterns in data. Finding patterns by operators would mean an extra overload to what they are already supposed to do. Because of abundance of digital data, oil and gas industry, and in particular refineries can benefit from what data mining or machine learning has to offer. Therefore there is an increasing need for data miners and use of data mining techniques. Mohaghegh (2005) shows the need for use of data mining and machine learning techniques to deduce information and knowledge from the data that are collected in the oil industry. The past decades have seen an increased use of data mining techniques across all branches of science and engineering. To see the diversity of applications of data mining refer to (Chang et al., 2010; Saybani et al., 2009; Assous et al., 2010; Bartok et al., 2010; Ehsan Hajizadeh et al., 2010; Selim Gullulu and Seker, 2011). Although there has been much research on time series models, however to the best of our knowledge, we know of no other work in which derived time series have been used. Objective of

*Corresponding author. E-mail: saybani@gmail.com, saybani@siswa.um.edu.my.

this paper is to use data mining techniques and fuzzy logic to detect and predict sensor faults in an oil refinery. Methods presented here show how to predict irregularities in the system. It estimates the time period needed for the maintenance of sensors.

Optimizing the time and increasing maximum output of the plant are other significant issues emerging from this research. Researchers of this paper were also motivated by a study carried out by Schwabacher et al. (2009). We realized that similar significance, conditions and goals exist at National Aeronautics and Space Administration (NASA) as it does in a refinery. Similarly, detecting faults in sensor data in a refinery is important for at least the following reasons:

a) It can be helpful in making crucial decisions such as whether or not to stop a production process, when crucial information is missing and before reaching a critical situation.
b) Predicting faults from recorded sensor data can help to determine what kind of maintenance is needed in the future.
c) Recurring faults in historical data covering a long period of time can produce values about the quality of sensors used; this can help to be selective when purchasing sensors.
d) The knowledge gained here could lead to improve design engineering of the refinery.

Currently, human experts try to detect sensor failures or anomalies; they watch and study the data during production process, but they have limited aid to check all sensor values. This approach is also very tedious and humans may fail to recognize faults that involve the relationships among large numbers of variables. A production delay is usually not desirable, therefore using an automated, in advance faults warning, it is a precious tool for the engineers and operators. Workloads of operators may easily let them not to detect the fault, since faults could happen too quickly for humans to detect them and react before they become in extreme case catastrophic. This research introduces a data mining technique to detect refinery sensor data anomaly and predict the occurrence of next faulty event. To cluster sensor data, a fuzzy-based predictor model was generated automatically using subtractive fuzzy clustering method. Derived time series, a new kind of time series is introduced and proposed. Furthermore, this paper shows two prediction models namely: auto regressive integrated moving average and autoregressive tree models which are used for predicting the next occurrence of sensor failure. And results will be compared. Models presented here can serve as an additional tool for engineers and operators to optimize the oil refining process. In the following study, we breifly touch the mathematical concepts behind methods used in this paper for predicting sensor faults or anomalies, first we start with definnition of time series, then auto-regressive integrated moving average (ARIMA) will be presented followed by autoregressive tree models (ARTxp).

Literature review, methodolyg and results are discussed in this study, respectively. Derived time series, a new concept introduced by the authors of this paper is introduced in the section of methodology. A time series is a chronological sequence of measurements on a particular variable that follow non-random orders. Usually the measurements are taken at equally spaced time intervals (days, months, years), however the sampling could be irregular. A time series analysis has two goals: 1) building a model that represents the nature of a time series, and 2) using the model to predict (forecast) future values of the time series. To achieve these goals it is required to establish a pattern and describe it. Then we can interpret and integrate it with other data. StatSoft.com (2010) and DTREG (2010) describes the characteristics of a time series so: the value of a time series with a regular pattern should be a function of previous values. Let us assume Y is the target value which the model wants to predict, and Y_t is the value of Y at time *t*, then the model could be written as follows:

$$Y_t = f(Y_{t-1}, Y_{t-2}, Y_{t-3}, \ldots, Y_{t-n}) + e_t \qquad (1)$$

Where Y_{t-1} is the value of Y for the previous observation, Y_{t-2} is the value two observations ago, etc., and e_t represents noise that does not follow a predictable pattern (this is called a random shock). Values of variables occurring prior to the current observation are called 'lag values'. If a time series follows a repeating pattern, then the value of Y_t is usually highly correlated with $Y_{t\text{-cycle}}$ where the cycle is the number of observations in the regular cycle. The goal of building a time series model is the same as the goal for other types of predictive models which is to create a model such that the error between the predicted value of the target variable and the actual value is as small as possible (DTREG, 2010). Usually modeling and predicting procedures involve knowledge about the mathematical model of the process. However, in normal life, very often the patterns of the data are not clear, data collections and observation have a lot of noise and errors, and therefore we still need not only to find the hidden patterns in the data. However, also try to generate forecasts. Box and Jenkins (1976) developed a popular methodology called ARIMA. It is powerful and flexible, but it is also complex and not easy to use. Box-Jenkins model assumes that the time series is stationary (NIST/SEMATECH, 2010). We briefly explain mathematical background of ARIMA here, for more mathematical details refer to (Box and Jenkins, 1994).

Autoregressive (AR) model is written as (NIST/SEMATECH, 2010):

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + A_t \qquad (2)$$

where X_t is the time series, A_t is the white noise, and:

$$\delta = (1 - \sum_{i=1}^{p} \phi_i)\mu \tag{3}$$

where µ is the process mean, p is called the order of AR model.

Moving average (MA) model is written as (NIST/SEMATECH, 2010):

$$X_t = \mu + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \cdots - \theta_q A_{t-q} \tag{4}$$

Where X_t is the time series, A_{t-i} is the white noise, µ is the mean of the series, $\theta_1, \ldots, \theta_q$ are the parameters of the model and q is called the order of MA model.

Putting AR and MR together we get the Box-Jenkins ARMA model which is written as (NIST/SEMATECH, 2010):

$$X_t = \delta + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots$$

$$+ \phi_p X_{t-p} + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \cdots - \theta_q A_{t-q} \tag{5}$$

Mathematically, ARIMA model is written as (James-Madison-University, 2010):

$$W_t = \mu + \frac{\theta(B)}{\phi(B)} a_t \tag{6}$$

Where t is index time and W_t is the response series Y_t or a difference of the response series.

μ is the mean term, B is the backshift operator, that is, $BX_i = X_{i-1}$,

B is the autoregressive operator, represented as a polynomial in the back shift operator:

$$B = 1 - \phi B - \ldots - \phi_p B^p \tag{7}$$

B is the moving-average operator represented as a polynomial in the back shift operator:

$$B = 1 - \boldsymbol{\theta} B - \ldots - \boldsymbol{\theta}_q B^q \tag{8}$$

a_t is the independent disturbance also called random error.

The general model includes autoregressive as well as moving average parameters, and particularly it includes differencing in the formulation of the model. The model ARIMA (p, d, q) has three types of parameters which are explained as follows:

i) p is the number of autoregressive parameters. It specify which previous values from the series are used to predict current values.
ii) d is the number of non-seasonal differences (or the number of differencing passes). If trends are present, differencing becomes necessary. The order of differencing corresponds to the degree of series trend for example first-order differencing specifies linear trends, second-order differencing accounts for quadratic trends, and so on.
iii) q is the number of lagged forecast errors in the prediction equation (or moving average parameters). Moving average orders tell how deviations from the series mean for previous values are used to predict current values. For example, moving-average orders of 1 and 2 specify that deviations from the mean value of the series from each of the last two time periods be considered when predicting current values of the series.

ARTxp algorithm was developed by Microsoft Research, it is based on the Microsoft Decision Trees algorithm which is an autoregressive tree (ART) model for representing periodic time series data. This algorithm relates a variable number of past items to each current item which is being predicted (Microsoft, 2008). The decision tree produces piecewise-linear AR model. The model uses Bayesian technique to learn the structure and parameters of the decision tree (Christopher et al., 2002). We briefly explain mathematical background of ART here, more mathematical details sre given by Christopher et al. (2002). The ART model for a tree with length p is written as:

$$f(y_t|y_{t-p}, \ldots, y_{t-1}, \theta) = \prod_{i=1}^{L} f_i(y_t|y_{t-p}, \ldots, y_{t-1}, \theta_i)^{\phi_i} = \prod_{i=1}^{L} N(m_i + \sum_{j=1}^{p} b_{ij} y_{t-j}, \sigma_i^2)^{\phi_i} \tag{9}$$

Where L is the number of leaves, $\theta = (\theta_1, \ldots, \theta_L)$ and $\theta_i = (m_i, b_{i1}, \ldots, b_{ip}, \sigma_i^2)$ are the model parameters for the linear regression at leaf l_i, i = 1, ..., L, $N(\mu, \sigma_i^2)$ is a normal distribution with mean µ and variance σ^2.

ARTxp model makes use of posterior probalbility defined by Bayes for learning and forecasting purposes. Posterior probablity is given by Murphy (2010):

$$posterior = \frac{liklihood * prior}{marginal\ likelihood}$$

or in symbols it is written as:

$$P(R = r|e) = \frac{P(e|R=r)P(R=r)}{P(e)} \tag{10}$$

Where $P(R = r|e)$ denotes the probability that random variable R has value r given evidence e. The denominator is called the marginal likelihood and gives the prior probability of the evidence. The likelihood of the ARTxp model is denoted by:

$$p\left(y_{p+1},\dots,y_T \middle| y_1,\dots,y_p,\theta,s\right) \\ = \prod_{t=p+1}^{T} f(x_{p+1}^t \; | x_1^t,\dots,x_p^t,\theta,s) \qquad (11)$$

This is the liklihood for an ordinary regression model with target variable X_{p+1} and regressor variables $X_1,\dots, X_p$.

LITERATURE REVIEW

Varun et al. (2009) mention that anomaly detection is about finding patterns in data which does not conform to expected behavior. They refer to these nonconforming patterns as anomalies. They also note that in the literature the same definition is also used for outliers, exception and novelties. Victoria and Austin (2004) borrow the following definition for outlier from Grubbs (1969): "An outlying observation or "outlier" is one that appears to deviate markedly from other members of the same sample in which it occurs." There is a difference between anomaly and novelty, and the difference is, that novel patterns are usually embedded into the normal model after their initial detection. In their survey, Varun et al. (2009) argue that anomaly detection has many applications and refer to Kumar (2005), where he brings a sample of anomalous traffic pattern in a computer network as an indication of unlawful transfer of sensitive data; or where Clay et al. (2001) talk about indication of malignant tumors, whether or not the MRI image shows anomalies; or where (Ryohei et al., 2005) mention that anomalous reading from a space craft sensor could be a hint to a fault somewhere in the space craft. Before performing any data analysis, noise removal is essential and according to Varun et al. (2009) and Rousseeuw and Leroy (1987) dealt with unwanted noise in the data. Also, Teng (1990) tackled the issue of unwanted noise. To define clear region of normal and abnormal data is usually very difficult, very often the borderline between this two regions is so fuzzy, that it makes difficult to say, to which region the data belong. Sometimes noise behaves like actual anomaly which makes it difficult to remove (Charu and Yu, 2001), Varun et al. (2009), Victoria and Austin (2004). According to Varun et al. (2009), the lack of enough abnormal data for the purpose of training or validating a model is a major issue. Anomaly could be understood differently for different applications, for example a small deviation of body temperature in medical domain is considered as anomaly, while similar deviation in a refinery process is considered as normal. Due to these difficulties, most of the existing anomaly detection models solve a specific formulation of a problem. Varun et al. (2009) mention that the formulation depends on various factors such as the nature of the data, its availability, types of anomalies to be detected and so on. Surveys on anomaly detection, review of articles and book reviews were conducted by many researchers such as Varun et al. (2009), Animesh and Park (2007), Bakar (2006), Malik et al. (2006), Victoria and Austin (2004), Markos and Singh (2003a) and Markos and Singh (2003b). For an extended survey on anomaly or outlier detection given by Victoria and Austin (2004) and Varun et al. (2009). Victoria and Austin (2004) bring some reasons why anomaly or outlier may occur; among them are mechanical failures, human error, instrument error, system error. In this sense, in this paper, sensor failure is considered as an instrument or a mechanical failure.

Varun et al. (2009) state that industrial units get damaged due to continuous usage and this should be detected as early as possible to prevent losses. Weilin et al. (2009) state that very often the life span of sensors depends on measurement frequency. Therefore it is essential to detect or predict sensor failure in advance. Varun et al. (2009) also mention that sensor networks have recently turned to an important research area, because collected sensor data have many unique characteristics. Anomalies that are detected through sensor data could be interpreted in many ways, it could be that one or more sensors are faulty or some components are faulty or something else is happening, thus it is important to study these phenomenon and characteristics. As Singh (2006) mentions, reducing equipment downtime, increasing reliability and availability of the equipments are considered as the most important strategical objectives , which can optimize the life cycle of the equipment. He considers costs associated with manufacturing design as fixed and predetermined, and therefore he suggests, in order to be competitive in the open market, the users have no other choice than optimizing life cycle of engines during their operation and maintenance. In the context of this paper, “engines” here corresponds to the equipments used in a refinery. A refinery needs to be operative all the time and it must function properly all the time with least amount of costs. Different researchers have used different techniques for anomaly detection in sensor networks. Varun et al. (2009) mention some of them in their survey and refer to Janakiram (2006) who used Bayesian networks, Joel et al. (2006) used rule-based systems, Phuong et al. (2006) used parametric statistical modeling, Subramaniam et al. (2006) as well as Kejia (2007) used nearest neighbor-based techniques and Daniela and Madden (2006) used time series techniques to forecast most likely future values. Schwabacher et al. (2009) mentions that model-based approach is one way of detecting anomalies; this approach encodes human knowledge into a model. But this model is very time consuming and labor intensive, and the feasibility of modeling every part of a complex system is very low. Therefore Schwabacher et al. (2009)

xItemTag	xValue	xDCSTime
03TC131_PV	438.0219	2009-12-18 11:38:46.627
03ZH2_PVFL	True	2009-12-18 11:38:46.767
03FC101_PV	214.8918	2009-12-18 11:38:46.407
03PC808_PV	-2.4614...	2009-12-18 11:38:46.407
03TC136_PV	513.5984	2009-12-18 11:38:46.407
03TC140_PV	344.2319	2009-12-18 11:38:46.407
03TI138_PV	371.0848	2009-12-18 11:38:46.407
03FC102_PV	54.97715	2009-12-18 11:38:46.000
03FC104_PV	54.97016	2009-12-18 11:38:46.000
03FC106_PV	54.90555	2009-12-18 11:38:46.000
03FC108_PV	50.76638	2009-12-18 11:38:46.000
03FI401_PV	16.92881	2009-12-18 11:38:46.000
03TI101_PV	308.8866	2009-12-18 11:38:46.000
03TI107_PV	440.8369	2009-12-18 11:38:46.000
03TI114_PV	443.3936	2009-12-18 11:38:46.000
03TI121_PV	439.3969	2009-12-18 11:38:46.000
03TI128_PV	435.2	2009-12-18 11:38:46.000
03TI137_PV	350.7699	2009-12-18 11:38:46.000
03TI203_PV	232.2138	2009-12-18 11:38:46.000

Figure 1. Typical data stream.

used data-driven approach, where anomalies are detected based on the data. This study however, shows a new way of predicting anomalies based on derived time series, applied on sensor data. Victoria and Austin (2004) conclude that "there is no single universally applicable or generic outlier detection approach." They recommend that developer should choose a suitable algorithm in a way that best fits his/her need. It should have correct distribution model, correct attribute type and it should be scalable. There are many researches dealing with time series, sensor and forecasting, but the search engine of ISI Web of Knowledge revealed that none of the available articles has dealt with forecasting sensor failure. Key words used for this search were “time series” + sensor + forecast.

The result of this search showed that French et al. (2010) were merely interested in estimating evapotranspiration. For the estimation they extrapolated remotely sensed inputs. In another study, the consistency of records derived from advanced very high resolution radiometer, SPOT-vegetation, SeaWiFS, moderate resolution imaging spectroradiometer and Landsat ETM+ was evaluated by Brown et al. (2006). In another research, Alexander et al. (1999) described a technique in which data from passive microwave sensors as well as infrared sensors and lightning hash observations together with digital image morphing were combined to yield a continuous time series of rain rates which may be assimilated into a mesoscale model. For validating the integrated water vapor from weather forecast models, Kopken (2001) used time series of vertically integrated water vapor derived from ground-based global positioning system sites in Sweden and Finland. In another research, in order to produce a spatially consistent estimate using the same set of inputs over all regions and times, Sapiano et al. (2008) studied a new gridded global analysis of precipitation using optimum interpolation based on the defense meteorological satellite program and the forecast precipitation from earlier re-analysis (Cohen et al., 2008; Cunha et al., 2010; Forzieri et al., 2010; Gibescu et al., 2009; Rodrigues and Gama, 2009; Singh et al., 2009; Wang et al., 2007). In contrast to these researches, we focused on forecasting the next occurrence of sensor failures. This paper presents a new algorithm to detect and predict sensor failures; however it is applicable in any situation where time series experience anomalies.

METHODOLOGY

At the beginning of every data mining, the process of data collection is very important. The process of getting data, which is used in this paer was discussed in details by Saybani and Wah (2010). Let us consider having a set of n sensors S = {s_1, s_2, ..., s_n} where n $\in N$ (natural numbers) and n≥ 1. Sensor data were gathered as a collection of data streams, data arrives as a string of values for a predefined n number of sensors in the form of: (s_1, v_1, t_1), (s_2, v_2, t_2), ..., (s_n, v_n, t_n), where (s_i, v_i, t_i) indicates the value of sensor s = i at time t_i. After each period α, a new collection of data is collected and saved. In general α = 5 and it means every 5 min. When reading the data from the database, the data comes as a data stream. The table in the database has basically 4 columns, however the first 3 columns were interesting for this research. The first column contains sensor id s_i the second column is the value v_i of sensor s_i at time t_i and the third column is the DCS system time t_i. Figure 1 shows a typical data stream for some sensors. For classification of whether or not, a sensor has experienced a faulty condition, this research used MATLAB’s Fuzzy Logic toolbox tools; in particular we used the genfis2 method to generate a model. This method uses subtractive fuzzy clustring algorithm which is fast for estimating the number of clusters and centroids in a set of data. The generated model is a Sugeno-type fuzzy inference system. It was introduced by Sugeno (1985), that is very similar to the method introduced by Mamdani and Assilian (1975) which was based on Lotfi Zadeh's 1973 paper on fuzzy algorithms for complex systems and decision processes (Zadeh, 1973). To handle each sensor’s data, a fuzzy clustering of sensor data was done. Each sensor is considered as a class for itself, therefore the model is capable of dealing with n classes. Maximum number of classes that were used for this research were intentionally limited to 18 for simplicity reasons. In MATLAB, we used the following commands to load data and generate clustering: load testing.txt; fismat = genfis2 (training_input, training_output, 0.5) genfis2 has the following syntax:

“fismat = genfis2(Xin,Xout,radii)

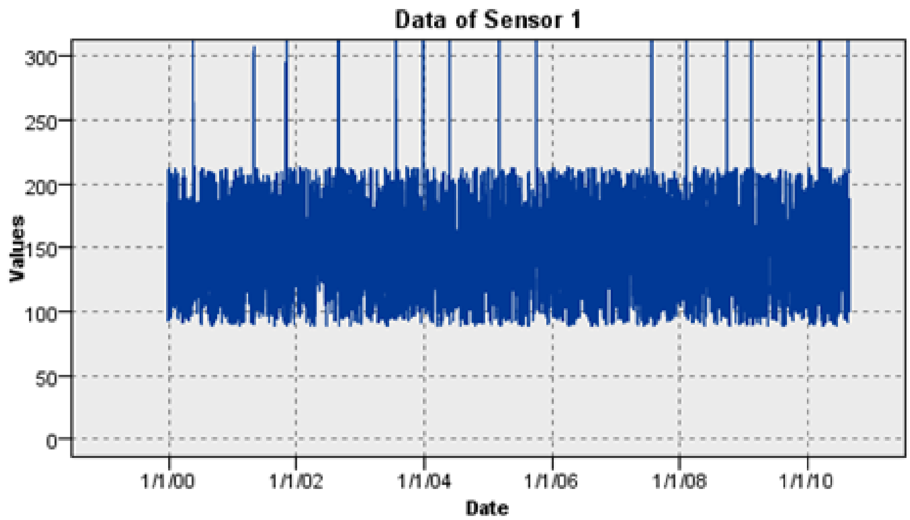

Figure 2. Data of a sensor over the time.

The arguments for genfis2 are as follows:

Xin is a matrix in which each row contains the input values of a data point.

i) Xout is a matrix in which each row contains the output values of a data point.
ii) Radii is a vector that specifies a cluster center's range of influence in each of the data dimensions assuming the data falls within a unit hyperbox" (MathWorks).

For simulation and evaluation or testing the model, we used MATLAB's evalfis method. We used the following commands: load training.txt; out = round [evalfis (testing_input, fismat)];
evalfis performs fuzzy inference calculations and has the following syntax: output = evalfis (input, fismat) and evalfis has the following arguments:

Input

A number or a matrix specifying input values. If input is an M x N matrix, where N is number of input variables, then evalfis takes each row of input as an input vector and returns the M x L matrix to the output variable, where each row is an output vector and L is the number of output variables.

Fismat

A FIS structure to be evaluated (MathWorks).

The result of clustering is in best case 0 or 1, where in our definition, 0 stands for clusters of normal data and 1 for clusters of data where sensor data indicate faultiness. If the determination of clustering becomes difficult, we can use adaptive-network-based fuzzy inference system (ANFIS). ANFIS uses a hybrid learning procedure which can construct an input-output mapping based on both human knowledge (in the form of fuzzy if-then rules) and stipulated input-output data pairs. In case of a simulation, the ANFIS architecture is employed to model nonlinear functions, identify nonlinear components on-line in a control system, and predict a chaotic time series, all yielding remarkable results (Jang, 1993). The significance of this classification lies in the following factors, first, this model can be used for online data streams, it has proven to be very efficient, and data streams coming from DCS or OPC Server have online nature, second, operators and engineers at the refinery can only monitor limited number of sensors at the same time, but there are thousands of sensors across the refinery, thus the model used in this research is seen as additional tool for warning the operators, when one or more sensors fail. This model is capable of identifying all faulty sensors. Next step in our model building was to visualize the data to see, whether or not, there are sensor failures, Figure 2 shows the data of a sensor over time. Figure 2 shows data behaviour of the sensor 1. For each sensor s_i (i = 1, 2, ..., n), there exists m features or properties which are shown in a property set P = {p_1, p_2, ..., p_m}. Major properties used in this model were p_1 = class id, p_2 = date and time of data recording, and p_3 = sensor value. After initial clustering, an automatic data cleansing on all classes were done. Missing data were replaced by mean values of each class and records with values outside of minimum and maximum range were discarded. These data were saved in a warehouse for further treatment. For data cleansing and dealing with missing data, a tree like decision algorithm was used to solve issues such as unusual values and missing values. Figure 3 illustrates a table with values of multi sensors. Usually time series are applied on existing data to predict their future values. In case of sensor data, time series would predict future common values for sensors and not the next occurrence of the sensor failure. To overcome this problem we had to create new time series from the existing time series for each sensor with a history of failures. We defined a time series which is created out of another time series as "derived time series" (DTS). To the best of our knowledge, this is the first time that such term is defined.

The steps to create a DTS were explained as follows: In our time series, values are usually numeric and represent temperature, pressure, flow and so on depending on sensor type. Value "NaN" indicates a fault of sensor. To measure the distribution of the data fault, it was necessary to calculate the time difference between each successive fault events, time difference and time of occurrence of each fault were saved in a separate table as shown in Figure 4. Each record has a structure of (t_i, Δt_i), where t_i is the time when the failure is recorded and time difference is given by $\Delta t_i = t_{i+1} - t_i$, where (i = 0, 1, 2, ...). In Figure 4, the first record might be biased because we did not know when exactly

Date	Time	Sensor1	Sensor2	Sensor3	Sensor4	Sensor5	Sensor6	Sensor7
2010-08-03	00:00:44.482	168.322	47.916	198.643	64.612	62.370	2627.341	54.858
2010-08-04	00:00:11.782	216.199	46.520	201.761	63.590	63.328	2714.166	55.891
2010-08-05	00:00:51.431	192.573	58.668	201.321	62.828	65.240	2655.257	55.992
2010-08-06	00:00:23.727	117.155	64.790	206.390	60.738	67.118	2675.125	57.234
2010-08-07	00:00:40.480	205.061	72.719	206.083	64.455	65.239	2629.300	58.295
2010-08-08	00:00:36.497	99.083	49.758	205.095	62.317	64.037	2628.358	57.035
2010-08-09	00:00:19.295	135.282	35.393	198.131	59.630	65.538	2692.695	56.625
2010-08-10	00:00:18.779	145.650	22.571	206.284	64.818	61.659	2641.380	56.036
2010-08-11	00:00:14.660	189.940	73.455	202.857	63.507	62.082	2657.797	55.422
2010-08-12	00:00:17.430	189.818	75.209	205.211	63.192	64.180	2712.368	57.168
2010-08-13	00:00:18.282	174.328	41.867	205.613	60.271	63.318	2627.992	54.733
2010-08-14	00:00:44.152	213.072	58.334	202.331	64.253	63.544	2647.577	54.274
2010-08-15	00:00:33.954	149.021	57.153	204.838	59.828	66.829	2651.540	55.084
2010-08-16	00:00:00.291	186.750	41.902	205.421	58.874	62.872	2668.648	54.304
2010-08-17	00:00:32.672	156.731	75.942	198.595	58.340	62.183	2671.207	54.222
2010-08-18	00:00:40.868	144.222	85.212	204.338	60.605	65.033	2626.145	58.244
2010-08-19	00:00:44.659	98.085	63.159	205.221	60.491	65.081	2674.012	57.665
2010-08-20	00:00:06.860	122.236	26.210	199.607	58.832	64.735	2632.611	55.565
2010-08-21	00:00:01.280	107.261	19.301	206.302	59.172	64.543	2646.764	56.528
2010-08-22	00:00:42.310	124.276	52.660	206.385	59.344	63.098	2706.634	55.865
2010-08-23	00:00:05.606	145.089	87.214	199.424	63.654	62.056	2696.614	55.531

Figure 3. Records of some sensors.

Date	Value
2000-05-24	145....
2001-05-07	346....
2001-11-07	182....
2002-08-31	294....
2003-07-28	328....
2003-12-30	153....
2004-05-29	150....
2005-03-09	282....
2005-10-02	206....
2006-06-15	252....
2006-11-10	144....
2007-07-29	259....
2008-02-12	195....
2008-10-02	230....
2009-02-17	137....
2009-10-06	229....
2010-03-16	158....
2010-08-26	159....

Figure 4. Time difference.

sensor 1, for example, was used for the first time; therefore we assumed that the time of first recording was the first time of usage. Hence, in the example shown in Figure 4, value of 145 indicates number of days that were elapsed from the start date t_0, until t_1 when the first failure was recorded.

$$(\Delta t_1 = t_1 - t_0 = 145).$$

DTS are useful when initial time series are not practical or meaningful for prediction, however their DTS can be used to predict irregularities in a system, or in ideal case and determine regularities. Authors of this paper see applications for DTS in various scientific and engineering areas such as in manufacturing, production and health industry to name a few. The next step after creating DTS table was to forecast the next occurrence of a sensor failure. For this authors of this paper used, time series model of SPSS PASW v.13 and v.14 (former clementie), which uses ARIMA algorithm to forecast values, we also used another data mining tool (forecast model) provided as an adds-in in Microsoft Excel 2007; it uses data mining tools of Microsoft SQL 2008 Server. Microsoft Forecast model uses a blend of ARTxp and ARIMA. Details on how this model works was given by Christopher et al. (2002) and Microsoft (2008). There are different algorithms that can do the forecasting such as time series regression model and exponential smoothing, however ARIMA has shown good forecasting performance (Sanchez, 2004), for this reason we decided to use ARIMA in our model. Pseudocode of the algorithm used in this paper is presented as follows:

For each sensor:

Verify anomaly,
IF anomaly then//report anomaly,
Perform DST,
Predict next anomaly occurrence,
Else,
Read next sensor value,
End IF,
End For.

			Results of two models					
Final Estimated values			Actual	Estimation value 1		Actual	Estimation of value 2	
Sensor Id	All-Data-PASW	All-Data-SQL	Values 1	PASW v.14	SQL 2008	Values 2	PASW v.14	SQL 2008
01	214	191	229	217	119	137	221	133
02	249	249	240	249	249	292	258	258
03	251	268	182	249	249	327	254	254
04	257	157	272	260	150	132	259	259
05	221	261	234	225	250	99	224	234
06	222	237	148	217	217	190	222	258
07	214	135	106	196	144	213	201	174
08	231	340	334	240	229	254	233	239
09	235	298	128	234	303	131	242	244
10	226	163	102	222	308	237	230	115
11	250	250	110	246	244	261	256	135
12	217	205	351	236	243	272	215	164
13	254	154	262	257	258	301	257	315
14	263	263	323	256	256	101	250	251
15	233	116	269	235	211	268	233	155
16	240	194	317	242	163	212	237	207
17	217	268	112	218	129	104	225	177
18	240	242	301	233	233	118	228	305

Figure 5. Results of two forecasting models.

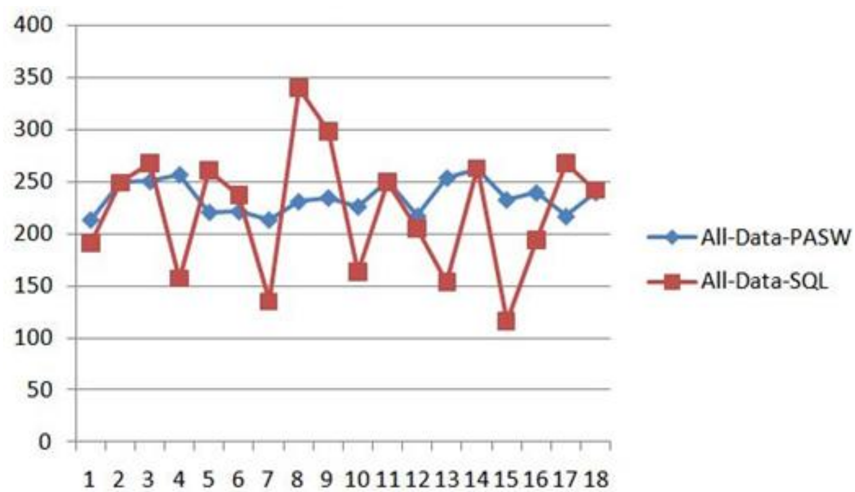

Figure 6. Graph of the final prediction.

RESULTS

For anomaly detection, this research used 5999 records for training and 5631 records for testing the model. This model achieved a classification accuracy of 100%. Models used in this paper are capable of detecting anomalies in a time series. Figure 5 illustrates the outcome of prediction occurance of sensor failures using two algorithms ARIMA in SPSS-PASW and Forcast in Microsoft Excel 2007 with a connection to Microsoft SQL Server 2008. Figure 6 shows the graph of the final prediction for the sensors, data sources for this graph come from columns 2 and 3 displayed in Figure 5. In order to see how well the aforementioned forecast models do, researchers decided to run the DTS data set against both models, but to exclude last record of each sensor. In other words, researchers wanted to know how well the methods can predict the last value. Figure 7 illustrate the result, data source for this graph are columns 5 and 6 (red line represents predicted values using PASW v.14's ARIMA model, green line represents predicted values using MS-SQL Server 2008's forecast model and column 4 represnts the values that were supposed to be predicted, represented by blue line – value 1. We ran prediction models against the DTS data set and this time, we omitted 2 values of each sensor from the data set. The purpose was to determine how well those models predict second last value of the data set with less available data set, graph of this test is shown in Figure 8. Data source for this graph comes from

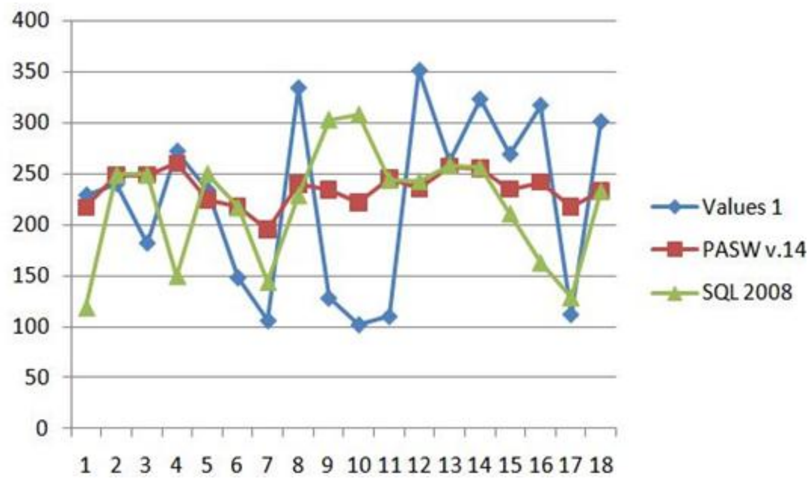

Figure 7. Graph of prediction for last value of DTS data set.

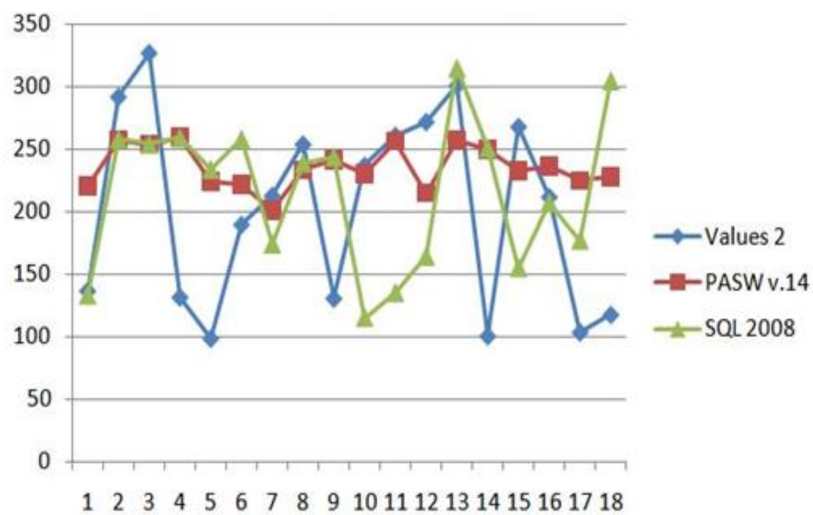

Figure 8. Graph of prediction for second last value of DTS data set.

the last 3 columns shown in Figure 5 (red line represents predicted values using PASW v.14's ARIMA model, green line represents predicted values using MS-SQL Server 2008's forecast model and column 4 represnts the values that were supposed to be predicted, represented by blue line - value 2.

CONCLUSIONS

In this research we have shown data mining techniques for classifying data streams at a refinery, a fuzzy-based predictor model was generated automatically using subtractive fuzzy clustering method; in particular we used fuzzy inference system and fuzzy clustering to cluster sensors. After clustering and identifying sensor failures, we created a new model for forecasting the occurance of next sensor failure. We determined time difference between each two consecutive sensor failures and the result was inserted in a DTS table as the input for the time series model. Researchers used Time Series Model in SPSS-PASW v.13 and v.14, as well as “forecast” model of Microsoft SQL Server 2008 add-ins for office 2007. Different models deliver different results, this is natural and is due to differences in algorithms used by these models. When it comes to decide what model should be used, we recommend , if there is a possibility of having multiple models, then in order to be on the safe side, one should use the value of such model that gives the lowest value, obviousley it is better to be prepared for it sooner than later. To the best of our knowledge, this research is the first of its kind carried out in a refinery, especially in Persian Gulf area and definitely in Iran. We are the creators of derived time series (DTS), and have shown that our model can be used to detect and predict sensor failures.

Our models can serve as additional tool, which could help engineers and operators to optimize the oil refinery productions. It gives experts at the oil refinery the opportunity to have two parallel models. They have the option to compare the models and their results and choose prefered model.

AKNOWLEDGEMENTS

Researchers thank Mr. Amin, Mr. Nasrabadi, Mr. Behroozi, Mr. Ghorbanipour, Mr. Khosravi, Mr. Kazemi, Mr. Masoomi, Mr. Bahrampour and Mr. Ghanizadeh for providing data and technical support.

REFERENCES

Alexander GD, Weinman JA, Karyampudi VM, Olson WS, Lee ACL (1999). The effect of assimilating rain rates derived from satellites and lightning on forecasts of the 1993 superstorm. Monthly Weather Rev., 127: 1433-1457.

Animesh P, Park JM (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. *Comput. Netw.,* 51: 3448-3470.

Hajizadeh E, Davari AH, Shahrabi J (2010). Application of data mining techniques in stock markets: A survey, J. Econ. Int. Finance 2(7): 109-118.

Bakar Z, Mohemad R, Ahmad A, Deris M (2006). A comparative study for outlier detection techniques in data mining. Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems. IEEE.

Bartok J, Habala O, Bednar P, Gazak M, Hluchý L (2010). Data mining and integration for predicting significant meteorological phenomena. *Procedia Comput. Sci.,* 1: 37-46.

Brown ME, Pinzon JE, Didan K, Morisette JT, Tucker CJ (2006). Evaluation of the consistency of long-term NDVI time series derived from AVHRR, SPOT-Vegetation, SeaWiFS, MODIS, and Landsat ETM+ sensors. *Ieee Trans. Geosci. Remote Sens .,* 44: 1787-1793.

Chang CD, Wang CC, Jiang CB (2010). Using Data Mining Techniques for Multi-Diseases Prediction Modeling of Hypertension and Hyperlipidemia by Common Risk Factors. Expert Systems with Applications, In Press, Accepted Manuscript.

Charu CA, Yu PS (2001). Outlier detection for high dimensional data. Proceedings of the 2001 ACM SIGMOD international conference on Management of data. Santa Barbara, California, United States: ACM.

Christopher M, David MC, Heckerman DY (2002). Autoregressive Tree Models for Time-Series Analysis. *In:* SIAM Int. Conf. Data Min. Soc. Ind. Appl Math., pp. 229-244.

Clay S, Lucas P, Sajda P (2001). Detection, Synthesis and Compression in Mammographic Image Analysis with a Hierarchical Image Probability Model. Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'01). IEEE Computer Society.

Cohen AB, Ravoori B, Murphy TE, ROY R (2008). Using Synchronization for Prediction of High-Dimensional Chaotic Dynamics. *Physical Review Letters,* 101.

Cunha M, Marcal ARS, Silva L (2010). Very early prediction of wine yield based on satellite data from Vegetation. *Int. J. Remot. Sens.,* 31: 3125-3142.

Daniela T, Madden S (2006). Time Series Forecasting For Approximate Query Answering In Sensor Networks, Springer Berlin-Heidelberg.

DTREG (2010). *Time Series Analysis* [Online]. Available: http://www.dtreg.com/TimeSeries.htm?gclid=CJnDr5Lr16QCFcZA6wodI0JjKQ [Accessed].

Forzieri G, Castelli F, Vivoni ER (2010). A Predictive Multidimensional Model for Vegetation Anomalies Derived From Remote-Sensing Observations. Ieee Trans. Geosci. Remot. Sens., 48: 1729-1741.

French AN, Hunsaker DJ, Clarke TR, Fitzgerald GJ, Pinter PJ (2010). Combining Remotely Sensed Data and Ground-Based Radiometers to Estimate Crop Cover and Surface Temperatures at Daily Time Steps. J. Irrigat. Drainage Eng. Asce*.,* 136: 232-239.

Box GEP, Jenkins GM (1994). Time Series Analysis: Forecasting and Control, Upper Saddle River, New Jersey, Perentice Hall.

Hajizadeh E, Davari AH, Shahrabi J (2010). Application of data mining techniques in stock markets: A survey, J. Econ. Int. Finance 2(7): 109-118.

Gibescu M, Brand AJ, Kling WL (2009). Estimation of Variability and Predictability of Large-scale Wind Energy in The Netherlands. *Wind Energy,* 12: 241-260.

Grubbs FE (1969). Procedures for Detecting Outlying Observations in Samples. *Technometrics,* 11**,** 1.

Teng H, Lu KCS (1990). Adaptive real-time anomaly detection using inductively generated sequential patterns. *In:* In Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy, IEEE Comput. Soc. Press., pp. 278-284.

James-Madison-University. 2010. *The ARIMA Procedure* [Online]. Available: http://www.jmu.edu/docs/sasdoc/sashtml/ets/chap7/sect8.htm [Accessed].

Janakiram D, Reddy V, Kumar A (2006). Outlier detection in wireless sensor networks using Bayesian belief networks. *In:* 1st Int. Conf. Commun. Syst. Software Middleware, pp.1-6.

Jang JSR (1993). ANFIS: adaptive-network-based fuzzy inference system. Syst. Man Cybernet., IEEE Trans*.,* 23: 665-685.

Joel B, Boleslaw S, Chris G, Ran W, Kargupta H (2006). In-Network Outlier Detection in Wireless Sensor Networks. Proceedings of the 26th IEEE Int. Conf. Distr. Comput. Syst. IEEE Comput. Soc.,

Kejia ZSS, Hong G, Jianzhong L (2007). Unsupervised Outlier Detection in Sensor Networks Using Aggregation Tree. *Adv. Data Min. Appl.,* 4632: 158-169.

Kopken C (2001). Validation of integrated water vapor from numerical models using ground-based GPS, SSM/I, and water vapor radiometer measurements. J. Appl. Meteorol., 40: 1105-1117.

Kumar V (2005). Parallel and Distributed Computing for Cybersecurity. *IEEE Distr.* Syst. Online., 6: 1.

Schwabacher M, Oza N, Matthews B (2009). Unsupervised Anomaly Detection for Liquid-Fueled Rocket Propulsion Health Monitoring*. J.* Aerospace Comput. Inf. Commun., 6: 464-482.

Malik A, Ken B, Alhajj R (2006). A comprehensive survey of numeric and symbolic outlier mining techniques. *Intell. Data Anal.,* 10: 521-538.

Mamdani EH, Assilian S (1975). An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man-Mach. Stud., 7: 1-13.

Markos M, Singh S (2003a). Novelty detection: a review- part 1: statistical approaches. *Signal Process.,* 83: 2481-2497.

Markos M, Singh S (2003b). Novelty detection: a review-part 2: neural network based approaches. Signal Process*.,* 83: 2499-2521.

MATHWORKS. *evalfis* [Online]. MathWorks. Available: http://www.mathworks.com/help/toolbox/fuzzy/evalfis.html [Accessed].

Mathworks. *genfis2* [Online]. MathWorks.com. Available: http://www.mathworks.com/help/toolbox/fuzzy/genfis2.html [Accessed 2010].

Microsoft. 2008. Microsoft Time Series Algorithm Technical Reference [Online]. Available: http://technet.microsoft.com/en-us/library/bb677216.aspx [Accessed].

Mohaghegh SD (2005). A new methodology for the identification of best practices in the oil and gas industry, using intelligent systems. J. Petrol. Sci. Eng., 49: 239-260.

Murphy, K. 2010. A brief introduction to Bayes' Rule [Online]. Available: http://www.cs.ubc.ca/~murphyk/Bayes/bayesrule.html [Accessed].

NIST/SEMATECH. 2010. e-Handbook of Statistical Methods [Online]. Available: http://www.itl.nist.gov/div898/handbook/ [Accessed].

Phuong T, Hung L, Cho S, Lee Y-K, Lee S (2006). An Anomaly Detection Algorithm for Detecting Attacks in Wireless Sensor

Networks.

Rousseeuw PJ, Leroy AM (1987). Robust regression and outlier detection, John Wiley & Sons, Inc.

Rodrigues PP, Gama J (2009). A system for analysis and prediction of electricity-load streams. *Intell. Data Anal.,* 13: 477-496.

Ryohei F, Yairi T, Machida K (2005). An approach to spacecraft anomaly detection problem using kernel feature space. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. Chicago, Illinois, USA: ACM.

Subramaniam S, Palpanas T, Papadopoulos D, Kalogeraki V, Gunopulos D (2006). Online outlier detection in sensor data using non-parametric models. *Proceedings of the 32nd international conference on Very large data bases.* Seoul, Korea: VLDB Endowment.

Sanchez J (2004) Introduction to Time Series A forecasting model for Absorbent Paper Towels. University of California (UCla).

Sapiano MRP, Smith TM, Arkin PA (2008). A new merged analysis of precipitation utilizing satellite and reanalysis data. J. Geophys. Res. Atmosp., 113.

Saybani MR, Teh Ying WAL (2009). Applied Data Mining Approach in Ubiquitous World of Air Transportation. In: Comput. Sci. Convergence Info. Technol., ICCIT '09. Fourth Int. Conf., pp. 1218-1222.

Saybani MR, Wah TY (2010). Data mining and data gathering in a refinery. Proceedings of the 10th WSEAS international conference on Applied computer science. Japan: World Scientific and Engineering Academy and Society (WSEAS).

Selim G, Seker S (2011). Signal based approach for data mining in fault detection of induction motor', Sci. Res. Essays 6(22): 4720-4731.

Singh M, Mishra VD, Thakur NK, Kulkarni AV (2009). Impact of climatic parameters on statistical stream flow sensitivity analysis for hydro power. J. Indian Soc. Remote Sens., 37: 601-614.

Singh SOAR (2006). Artificial Neural Networks in Fault Diagnosis- A Gas Turbine Scenario.

Statsoft.COM. 2010. *Time Series Analysis* [Online]. Available: http://www.statsoft.com/textbook/time-series-analysis/ [Accessed 2010].

Sugeno M (1985). Industrial applications of fuzzy control, Elsevier Science Pub. Co.

Varun C, Arindam B, Kumar V (2009). Anomaly Detection: A Survey. ACM Computing Surveys (CSUR), 41.

Victoria H, Austin J (2004). A Survey of Outlier Detection Methodologies. Artif. Intell. Rev., 22, 85-126.

Wang X, Ma JJ, Wang S, Bi DW (2007). Time series forecasting for energy-efficient organization of wireless sensor networks. Sensors, 7: 1766-1792.

Weilin W, Lafortune S, Girard AR, Feng L (2009). Dynamic sensor activation for event diagnosis. *In:* American Control Conference, 2009. ACC '09., pp. 4753-4758.

Zadeh LA (1973). Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans. Syst. Man Cybernet., 3: 28-44.