BRAIN. Broad Research in Artificial Intelligence and Neuroscience,
ISSN 2067-3957, Volume 1, July 2010,
Special Issue on Complexity in Sciences and Artificial Intelligence,
Eds. B. Iantovics, D. Rădoiu, M. Măruşteri and M. Dehmer

DATA MINING LEARNING MODELS AND ALGORITHMS ON A SCADA SYSTEM DATA REPOSITORY

Maria Muntean, Ioan Ileană, Corina Rotar, Mircea Rîşteiu

Abstract. This paper presents three data mining techniques applied on a SCADA system data repository: Naïve Bayes, k-Nearest Neighbor and Decision Trees.

A conclusion that k-Nearest Neighbor is a suitable method to classify the large amount of data considered is made finally according to the mining result and its reasonable explanation.

The experiments are built on the training data set and evaluated using the new test set with machine learning tool WEKA.

Keywords: *Data Mining, SCADA System Data.*

1. Introduction

Knowledge discovery in databases (KDD) represents the overall process of converting raw data into useful information. According to the definition given in [1], KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. This process consists of a series of transformation steps, from data preprocessing to post-processing of data mining results.

Data mining, the central activity in the process of knowledge discovery in databases, is concerned with finding patterns in data. It consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data [1].

Classification is one of the primary tasks in data mining. It represents the task of learning a target function (classification model) that maps each attribute set to one of the predefined class labels [2]. In other words it consists in assigning objects to one of several predefined categories.

The evaluation of the performance of a classifier is a complex process. The inducer's complexity, cost, usefulness, generalization error and success rate should be taken in consideration when evaluating the predictive performance for the learned model. The most well-known performance metric is the success rate, which is based on counting the test records correctly and incorrectly predicted by the classification model. These counts can be displayed as a two-dimensional confusion matrix, with a row and column for each class.

The most important examples of classifiers from literature are: Decision Tress, Naïve Bayes, Neural Networks, Association Rules, k-Nearest Neighbor and Support Vector Machines. For solving our problem we chosen three different classifiers: Naïve Bayes, k-Nearest Neighbor and Decision Trees.

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions.

Depending on the precise nature of the probability model, Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In spite of their naive design and apparently over-simplified assumptions, Naive Bayes classifiers often work much better in many complex real-world situations than one might expect [3].

An advantage of the Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the vari-

ables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Instance-based (IB) learning methods simply store the training examples and postpone the generalization (building a model) until a new instance must be classified or prediction made. (This explains another name for IB methods – lazy learning – since these methods delay processing until a new instance must be classified).

K-nearest neighbor, an IB learning method, is a supervised learning algorithm where the result of new instance query is classified based on majority of K-nearest neighbor category. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory. Given a query point, we find K number of objects or (training points) closest to the query point.

K-nearest neighbor (k-NN) method assumes all instances correspond to points in the n dimensional space. The nearest neighbors of an instance are defined in terms of the standard Euclidean distance.

Decision tree learning represents one of the simplest, yet most popular methods for inductive inference. It has been successfully applied to a wide variety of problems from medical diagnosis to air traffic control or the assessment of credit risk for loan applicants. Its popularity is justified by the fact that it has some key advantages over other inductive methods. First of all, decision trees offer a structured representation of knowledge (as disjunction of conjunctive rules). As a direct consequence, decision trees may be rewritten as a set of " if-then" rules, increasing human readability. Secondly, decision trees are robust to errors, requiring little or no data preprocessing. Other important features include the capacity of handling both nominal and numeric attributes, as well as missing values and a good time complexity even for large data sets.

Structurally, a decision tree is a graph, whose inner nodes are " branching nodes" , because they contain some attribute test; the leaves contain the classification of the instance; the branches of the tree represent attribute values. The tree classifies an instance by filtering it down through the tests at the inner nodes, until the instance reaches a leaf.

The technique employed for building a decision tree is that of top-down induction, which performs a greedy search in the space of possible solutions. The first decision tree algorithm was introduced by J.R.Quinlan in 1986, and was called ID3. A large proportion of the decision tree learners that have been

developed since are improved variants of this core method; the most successful of them was the C4.5 algorithm, also developed by Quinlan [4].

2. Data Analysis

We chose the well-known Weka environment as the data mining tool to implement the experiment. Originally proposed for didactic purposes, Weka is a framework for the implementation and deployment of data mining methods. It is also an open-source software developed in Java, released under the GNU General Public License (GPL), being currently available to Windows, MAC OS and Linux platforms [7]. Weka contains tools for classification, regression, clustering, association rules, data visualization and works with .arff files (Attribute Relation File Format) and also with files in .csv format (Comma Separated Values).

The classifiers are the most valuable resource that Weka provides, and can be used in a variety of ways, such as applying a learning method to a dataset and analyzing its output to learn more about the data; or using learned models to generate predictions on new instances; a third is to apply several different learners and compare their performance in order to choose one for prediction.

We chose three datasets that contains the temperature values in June of 2007 to implement the experiment. A large amount of temperature variation information, obtained by the data collection equipment, was recorded and accumulated in the database of the SCADA system used.

The datasets used are presented in Table 1.

Dataset	Number of instances
1-10.06.07	14403
11-20.06.07	14400
21-30.06.07	14397

Table 1. The datasets used in the experiment

Supervisory Control and Data Acquisition (SCADA) systems provide automated control and remote human monitoring of real world processes in many fields as: food, beverage, water treatment, oil and gas, utilities.

The SCADA system is used to monitor and control a plant or equipment and is a combination of telemetry and data acquisition. Data acquisition deals with the methods used to access information or data from the controlled

equipment while telemetry is a technique used in transmitting and receiving this information over a medium.

SCADA has traditionally meant a window into the process of a plant and/or a method of gathering of data from devices in the field. Today, the focus is on integrating this process data into the actual business, and using it in real time. In addition to this, today's emphasis is on using Open Standards, such as communication protocols (e.g. IEC 60870, DNP3 and TCP/IP) and 'off-the-shelf' hardware and software, as well as focusing on keeping the costs down [6].

Concerning SCADA systems, there are at least two main issues: the reliability of the system and the optimal management of the huge amount of data being transferred to the SCADA server by the communication systems [7].

Our paper deals with the second issue and intends to contribute to a better using of communication lines ant to an economy of storing space. One can ask: all the time, all the acquired data are of the same importance for the plant control? Maybe a preprocessing at sensor level and some decisions taken at this level are better solutions than passing all the data to the server.

2.1. Data Preparation

The original data set included noisy, missing and inconsistent data. Data preprocessing improved the quality of the data and facilitated efficient data mining tasks.

Before the experiment, we prepared data suitable to next operation as following steps:

- Delete or replace missing values;
- Delete redundant properties (columns);
- Data Transformation;
- Data Discretization;
- Export data to a required .arff or .csv format file [11].

The original and modified formats of data set are shown in Figure 1 and Figure 2.

Data visualization is also a very useful technique because it helps to determine the difficulty of the learning problem. We visualized with Weka single attributes (1-d) and pairs of attributes (2-d). The figure 3 shows the variation of the temperature in time.

	A	B	C
1	DateTime	Feedwater Temp Inlet to Ecom	
2	2007.06.01 00:00:00	156.139328	
3	2007.06.01 00:01:00	156.6922302	
4	2007.06.01 00:02:00	156.6774292	
5	2007.06.01 00:03:00	155.9240112	
6	2007.06.01 00:04:00	155.1660919	
7	2007.06.01 00:05:00	154.868927	
8	2007.06.01 00:06:00	154.7238159	
9	2007.06.01 00:07:00	153.9180756	
10	2007.06.01 00:08:00	153.3471527	
11	2007.06.01 00:09:00	152.4390717	
12	2007.06.01 00:10:00	152.0284729	
13	2007.06.01 00:11:00	152.0105743	
14	2007.06.01 00:12:00	151.9789734	
15	2007.06.01 00:13:00	151.4999084	
16	2007.06.01 00:14:00	150.9686279	
17	2007.06.01 00:15:00	150.81604	
18	2007.06.01 00:16:00	150.8099365	
19	2007.06.01 00:17:00	150.803833	
20	2007.06.01 00:18:00	150.8164063	
21	2007.06.01 00:19:00	151.4267426	
22	2007.06.01 00:20:00	151.7000732	
23	2007.06.01 00:21:00	150.9064636	
24	2007.06.01 00:22:00	149.9548187	
25	2007.06.01 00:23:00	149.2224731	
26	2007.06.01 00:24:00	149.4431915	
27	2007.06.01 00:25:00	150.0674286	

Sheet1 Sheet2 Sheet3

Select destination and press ENTER or choose Paste

Figure 1. Original Data Format

```
@RELATION date234
@ATTRIBUTE DateTime DATE "yyyy-MM-dd HH:mm:ss"
@ATTRIBUTE FeedwaterTempInlettoEcom NUMERIC
@DATA "2007-06-01 00:00:00",156.139328
"2007-06-01 00:01:00",156.6922302
"2007-06-01 00:02:00",156.6774292
"2007-06-01 00:03:00",155.9240112
"2007-06-01 00:04:00",155.1660919
"2007-06-01 00:05:00",154.868927
"2007-06-01 00:06:00",154.7238159
"2007-06-01 00:07:00",153.9180756
"2007-06-01 00:08:00",153.3471527
"2007-06-01 00:09:00",152.4390717
"2007-06-01 00:10:00",152.0284729
"2007-06-01 00:11:00",152.0105743
"2007-06-01 00:12:00",151.9789734
"2007-06-01 00:13:00",151.4999084
"2007-06-01 00:14:00",150.9686279
"2007-06-01 00:15:00",150.81604
"2007-06-01 00:16:00",150.8099365
"2007-06-01 00:17:00",150.803833
"2007-06-01 00:18:00",150.8164063
"2007-06-01 00:19:00",151.4267426
"2007-06-01 00:20:00",151.7000732
"2007-06-01 00:21:00",150.9064636
"2007-06-01 00:22:00",149.9548187
"2007-06-01 00:23:00",149.2224731
"2007-06-01 00:24:00",149.4431915
"2007-06-01 00:25:00",150.0674286
"2007-06-01 00:26:00",150.6728973
"2007-06-01 00:27:00",151.2347107
"2007-06-01 00:28:00",150.4357452
"2007-06-01 00:29:00",151.5431519
"2007-06-01 00:30:00",153.002243
"2007-06-01 00:31:00",154.0943604
```

Figure 2. Modified Data Format

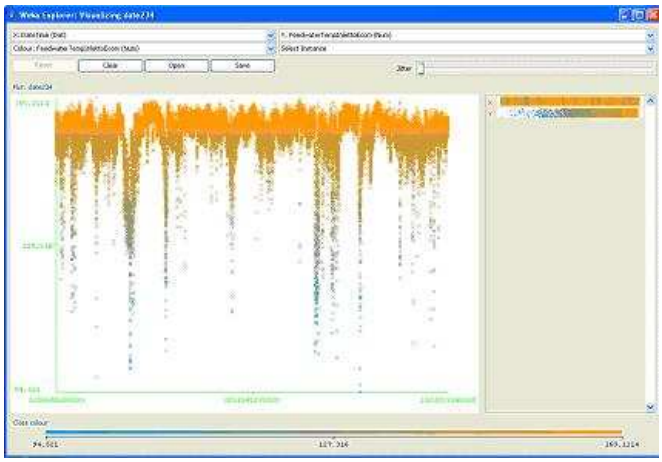

Figure 3. Data visualization

2.2.Data mining and interpretation of the results

A classification method was applied to assemble similar data points and to predict numeric quantities. In particular, we attempted to discover useful information and rules correlated to temperature values of the system in order to discard what could be regarded as irrelevant.

Based on the proposed framework, we chose *Naïve Bayes*, *kNN* and *J48* algorithms to implement classification. We tried to obtain clear results by choosing a 20% split percentage, which means that about 20% records were used as test data in the pre-implemented training process before classification [11]. The classifiers will be evaluated on how well they predicted the percentage of the data held out for testing. We want to determine which classifier is suited for our data set.

We concluded that *J48 Tree Classifier* model has a higher level of classification accuracy than the *Naïve Bayes Classifier* model, but the *IBk* algorithm is more adequate to our data set. The final results of the classification techniques are presented in the table below:

Dataset	**Classification accuracy**		
	Naïve Bayes Classifier	IBk Classifier	J48 Tree Classifier
1-10.06.07	98,72%	99,93%	99,93%
11-20.06.07	99,44%	100%	100%
21-30.06.07	99,06%	100%	99,96%
AVERAGE	**99,07%**	**99,98%**	**99,96%**

Table 2. The accuracy of the classification methods

3. Conclusions and future works

Classifier performance evaluation is an important stage in developing data mining techniques.

Our goal was to find the classifier that is suitable to the data set provided by SCADA system. The highest level of accuracy was matched in *IBk Classifier*. The three classes obtained after running the model allows a better optimization of the transmitted data traffic and of the necessary data storing space and projects the large amount of data to a lower dimensional space.

On the data acquisition system level we can program the transmission of warning and anomaly values and discarding normal values.

A future approach consists in a high sampling rate of data transmission from the three classes.

We also propose to study other SCADA parameters (e.g. pressure) and to develop one program that makes difference between acquisition system level and local storage of the functioning modes.

References

[1] J. Quinlan. Boosting first-order learning. Proceedings of the 7th International Workshop on Algorithmic Learning Theory, 1160:143–155, 1996.

[2] C. Nadal, R. Legault, and C. Y. Suen. Complementary algorithms for the recognition of totally unconstrained handwritten numerals. In Proceedings of the 10th International Conference on Pattern Recognition, volume A, pages 434–449, June 1990.

[3] Hand, DJ, & Yu, K. (2001). "Idiot's Bayes - not so stupid after all¿' International Statistical Review. Vol 69 part 3, pages 385-399. ISSN 0306-7734.

[4] J. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[5] http://www.dayton-knight.com/Projects/SCADA/scada_explained.htm

[6] http://www.bin95.com/certificate_program_online/control-systems-technology.htm

[7] I. Stoian, T. Sanislav, D. Căpăţînă, L. Miclea, H. Vălean, S. Enyedi, "Multi-agent and Intelligent Agents' Techniques Implemented in a Control and Supervisory Telematic System" , 2006 IEEE International Conference on Automation, Quality and Testing, Cluj-Napoca, 25-28 May 2006, pp. 463-468.

[8] E. K. Cetinkaya, "Reliability analysis of SCADA Systems used in the offshore oil and gas industry" , 2001.

[9] S. Wang, "Research on a New Effective Data Mining Method Based on Neural Networks" , 2008 International Symposium on Electronic Commerce and Security, Guangzhou City, 3-5 Aug. 2008, pp.195-198.

[10] B. Zheng, J. Chen, S. Xia, Y. Jin, "Data Analysis of Vessel Traffic Flow Using Clustering Algorithms" , 2008 International Conference on Intelligent Computation Technology and Automation, pp. 243-246.

[11] G. Wang, C. Zhang, L. Huang, "A Study of Classification Algorithm for Data Mining Based on Hybrid Intelligent Systems" , Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 371-375.

[12] P. Du, X. Ding, "The Application of Decision Tree in Gender Classification" , 2008 Congress on Image and Signal Processing, pp. 657-660.

[13] S. B. Shamsuddin, M. E. Woodward, "Applying Knowledge Discovery in Database Techniques in Modeling Packet Header Anomaly Intrusion Detection Systems" , Journal of Software, vol.3, no. 9, December 2008, pp. 68-76.

[14] Zengchang Qin, "Naive Bayes Classification Given Probability Estimation Trees" , Proceedings of the 5th International Conference on Machine Learning and Applications, 2006.

Maria Muntean, Ioan Ileană, Corina Rotar, Mircea Rîşteiu
Department of Mathematics and Informatics
"1 Decembrie 1918" University of Alba Iulia
email: *maria_munt2006@yahoo.com*