# KDS for Sericulture Cocoon Production

Kavita Bhosle
Marathwada Institute of Technology
Aurangabad(MS), India

## ABSTRACT
In this paper, we describe the formatting guidelines for ACM SIG Proceedings. Knowledge discovery is the process of analysing data for future planning. The nature of data and anomalies are different in different record data sets. The problem of detecting contextual anomalies in data sets is also different from the traditional anomaly detection problem. A time series is a chronological sequence of observations on a particular variable. Usually the observations are taken at regular intervals (days, months, years), but the sampling could be irregular. A time series analysis consists of two steps- building a model that represents a time series, and using the model to predict (forecast) future values. Anomaly depends on many factors such as temperature (season), Soil type etc. In our paper we proposed semi supervised learning for sericulture database. In sericulture the production of cocoons are analysed. The best, average and poor class label values depend on two approaches (1) Attributes like process knowledge, temperature, soil type, variety of mulberry plantation, use of fertilizers, turn of plantation etc. (2) Attribute contribution in terms of percentage. In semi supervised learning method in which input is line data stream and huge in nature. In the data stream, we can define class label val0ues for some instances but few instances are outlier. Outlier instances can be analysed using unsupervised learning. For semi supervised learning we are implementing Bayesian classification and rule based algorithm.

## 1. INTRODUCTION
Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends.

Anomalies are patterns in data that do not conform to a well defined notion of normal behaviour. Anomaly detection is related to, but distinct from noise removal [1] and noise accommodation [2], both of which deal with unwanted noise in the data. Noise can be defined as a phenomenon in data, which is not of interest to the analyst, but acts as a hindrance to data analysis. Noise removal is driven by the need to remove the unwanted objects before any data analysis is performed on the data. Input data can also be categorized based on the relationship present among data instances [3]. Most of the existing anomaly detection techniques deal with record data (or point data), in which no relationship is assumed among the data instances. In general, data instances can be related to each other. One of the examples is sequence data. In sequence data, the data instances are linearly ordered, e.g., time-series data, genome sequences, protein sequences such relationship among data instances become relevant for anomaly detection.

## 2. SEMI SUPERVISED ANOMALY DETECTION
Semi-supervised learning (SSL) is halfway between supervised and unsupervised learning. Semi-supervised learning will be most useful whenever there are far more unlabeled data than labeled. This is likely to occur if obtaining data points is cheap, but obtaining the labels costs a lot of time, effort, or money. The labels associated with a data instance denote if that instance is normal or anomalous. It should be noted that obtaining labeled data which is accurate as well as representative of all types of behaviors, is often difficult for on line data stream. Labeling is often done manually by a human expert and hence requires substantial effort to obtain the labeled training data set. Typically, getting a labeled set of anomalous data instances which cover all possible type of anomalous behavior is more difficult than getting labels for normal behavior therefore Supervised learning is not useful here. Moreover, the anomalous behavior is often dynamic in nature, e.g., new types of anomalies might arise, for which there is no labeled training data. In certain cases, such as air traffic safety, anomalous instances would translate to catastrophic events, and hence will be very rare. Based on the extent to which the labels are available, Semi Supervised Anomaly Detection techniques can operate better.

Suppose we knew that the points of each class tended to form a cluster. Then the unlabeled data could aid in finding the boundary of each cluster more accurately: one could run a clustering algorithm and use the labeled points to assign a class to each cluster.

The first one is *unsupervised learning*. Let X unsupervised = (x1; : : : ; xn) be a set of n examples (or learning points), where xi 2 X for all i 2 [n] := f1; : : : ; ng. Typically it is assumed that the points are drawn i.i.d. (independently and identically distributed) from a common distribution on X. It is often convenient to define the (n_d)-matrix X = (x>i )>i 2[n] that contains the data points as its rows. The goal of unsupervised learning is to find interesting structure in the data X. It has been argued that the problem of unsupervised learning is fundamentally that of estimating a density which is likely to have generatedX. However, there are also weaker forms of unsupervised learning, such as quantile estimation, clustering, outlier detection, and dimensionality

reduction.

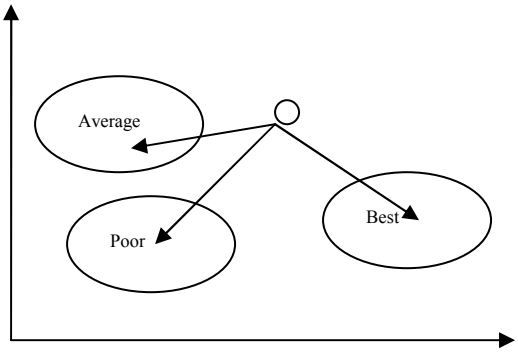


**Figure 1: Semi supervised learning**

## 3. CONTEXTUAL ANOMALY

If a time series has a regular pattern, then a value of the series should be a function of previous values. If Y is the target value that we are trying to model and predict, and $Y_t$ is the value of Y at time *t*, then the goal is to create a model of the form:

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

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 $Y_{t-cycle}$ where *cycle* is the number of observations in the regular cycle. For example, monthly observations with an annual cycle often can be modeled by

$$Y_t = f(Y_{t-12})$$

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. The primary difference between time series models and other types of models is that lag values of the target variable are used as predictor variables, whereas traditional models use other variables as predictors, and the concept of a lag value doesn't apply because the observations don't represent a chronological sequence.

We proposed semi supervised learning method in which input is on line data stream and huge in nature. In the data stream, we can define class label values for some instances but few instances are outlier or anomalies. The instances for which we can define class label values, we are using Bayesian Classification method.

For instances with class label values, construct classifier model using Bayesian Classification method. We are using one-class classification based anomaly detection technique and consider one class as one cluster. Assume that all training instances have only one class label. Such techniques learn a discriminative boundary around the normal instances using a one-class classification algorithm, e.g., one-class SVMs [8], one-class Kernel Fisher Discriminants [9]. Any test instance that does not fall within the learnt boundary is declared as anomalous or outlier. Once Normal instances are clustered using above method, considering all other instances as anomaly, we are applying rule based algorithm to detect abnormal cases.

### *Classification:*

This first step of the classification process can also be viewed as the learning of a mapping or function, $y = f(\boldsymbol{X})$, that can predict the associated class label *y* of a given tuple ***X***.

1. We used bays classification method. It is based on Bayes' Theorem of conditional probability. It is a statistical classifier: performs *probabilistic prediction, i.e.,* predicts class membership probabilities. A simple Bayesian classifier, *naïve Bayesian classifier*, assumes that different attribute values are independent which simplifies computational process. It has comparable performance with decision tree and selected neural network classifiers.

Let X be a data tuple ("evidence"): described by values of its n attributes

Let H be a hypothesis that X belongs to class C

Classification is to determine P(H|X), the probability that the hypothesis holds given the observed data sample X

Probability that X belongs to class C having known the attribute description of X

Given that X is instance. The class label value of X will be normal or abnormal or outlier. In any anomaly detection, class label value can be report using score or using labels. Scoring techniques assign an anomaly score to each instance in the test data depending on the degree to which that instance is considered an anomaly. Labeling techniques assigning a label (normal or anomalous or outlier) to each test instance. We used labeling technique for known instances.

Baye's theorem relates all these probabilities

Posteriori= Likelihood x priori / evidence

2. In second phase consider normal instance as three cluster like average, best and poor. We are using basian classification for above three category classification and anomaly or abnormal data can be classified using clustering. shown in figure 1

### *Rule based method-*

Rule based anomaly detection techniques learn rules that capture the normal behaviour of a system. A test instance that is not covered by any such rule is considered as an anomaly. Rule based techniques have been applied in multi-class as well as one-class setting.

| Varity of mulberry | Application of appropriate fertilizers dose (N-P-K) | Application of Manure | Irrigation at proper time | Soil type | Inter cultivation | Age of the leaf | Plant protection (insecticide or fungicide used?) | Mulberry plantation (Quality of leaf produced) |
|---|---|---|---|---|---|---|---|---|
| V-1 | 300-40-40 | 20 tons/hq | 12M-IR | Heavy | Regular | <65 days | Yes | Best |
| M-5 | 100-20-20 | 10 tons/hq | 8M-IR | Medium | Irregular | 90 days | No | Average |
| M-5 | 0-0-0 | 0 tons/hq | N-IR | Light | Not done | 120 days | No | Poor |
| V-1 | 100-20-20 | 10 tons/hq | 12M-IR | Medium | Irregular | 40 days | Yes | Best |
| M-5 | 300-40-40 | 20 tons/hq | 8M-IR | Medium | Regular | 65 days | No | Average |
| V-1 | 0-0-0 | 0 tons/hq | N-IR | Light | Regular | 60 days | No | Poor |
| M-5 | 300-40-40 | 10 tons/hq | 8M-IR | Medium | Regular | 50 days | Yes | Best |

**Table 1: Mulberry plantation (contributes 40-45 %)**

| DFLS | Environment | Type of rearing shade | Rearing skills and knowledge | Inputs and Environment |
|---|---|---|---|---|
| 603 | Summer | Brick wall | Yes | Average |
| 779 | Winter | Medium | No | Average |
| 1150 | Rainy | Simple | No | Best |
| 666 | Winter | Medium | Yes | Best |
| 776 | Rainy | Simple | No | Poor |
| 987 | Summer | Simple | Yes | Poor |
| 111 | Rainy | Medium | No | Average |

**Table 2: Inputs and Environment (contributes 35-40 %)**

| Disinfection and hygiene used | Bed disinfectant | Protection from disease |
|---|---|---|
| Yes | Yes | Best |
| Yes | No | Average |
| No | Yes | Average |
| No | No | Poor |

**Table 3: Protection from disease (contributes 10-15 %)**

| Mulberry plantation (contributes 40-45 %) | Inputs and Environment (contributes 35-40 %) | Protection from disease (contributes 10-15 %) | Cocoon Production |
|---|---|---|---|
| Best | Average | Poor | Best |
| Best | Poor | Average | Average |
| Best | Best | Best | Best |
| Best | Best | Average | Best |
| Best | Best | Poor | Best |
| Best | Best | Best | Best |
| Best | Average | Best | Best |
| Best | Poor | Best | Average |
| Average | Best | Poor | Best |
| Poor | Best | Average | Average |
| Average | Best | Best | Best |
| Poor | Best | Best | Average |
| Average | Poor | Best | Poor |
| Poor | Average | Best | Poor |

**Table 4: Cocoon Production dependability attributes and contributes**

As shown in figure consider cluster for normal class formed using supervised learning eg Bayesian classification method, Outside point may be anomaly or outlier. Now our aim is find this sequence pattern using association rule.

Association rule mining [10] has been used for one-class anomaly detection by generating rules from the data in an unsupervised fashion. Association rules are generated from a categorical data set. To ensure that the rules correspond to strong patterns, a support threshold is used to prune out rules with low support [3]. Association rule mining based techniques have been used for network anomaly detection [13], system call anomaly detection. An Apriori-Based Data Format Sequential Pattern Mining Algorithm can be used to identify sequential pattern.

All nonempty subsets of a frequent sequences must also be frequent- Large sequences are downward closed If we know that an sequences is small, we need not consider supersets of it as candidates because they also will be small. Apriori employs an iterative approach known as level-wise search, where k-sequences are used to explore k+1-itemsets.

# 4. CASE STUDY

In sericulture, Cocoon production depends on various factors (Table 1) like

1. Mulberry plantation (contributes 40-45 %)
    a. Varity of mulberry
    b. Application of appropriate fertilizers
    c. Application of Manure
    d. Irrigation at proper time
    e. Soil type
    f. Inter cultivation
    g. Age of the leaf
    h. Plant protection
2. Inputs and Environment (contributes 35-40 %)
    a. DFLS (disease free laying)
    b. Environment
    c. Type of rearing shade- brick wall, medium and simple rearing house
    d. Rearing skills and knowledge
3. Protection from disease (contributes 10-15 %)
    a. Disinfection and hygiene used
    b. Bed disinfectant-quantity and time

Class label represent cocoon production is average, best or poor as shown in table and final cocoon production depends on above main tree factors- Mulberry plantation, Inputs and environment and protection from disease. Observation – supervised or unsupervised learning are not giving accurate result for new instance individually. Instead of this if we used semi supervised learning then can minimize error. All experiments are performed using open source mining tool- WEKA.

# 5. CONCLUSION AND FUTURE SCOPE

In this case, ADT is more efficient than bays algorithm.

Key challenges faced by techniques in this category are:

1. The sequences might not be of equal length.

2. The test sequences may not be aligned with each other or with normal sequences.

For example, the first event in one sequence might correspond to the third event in another sequence. Comparing such sequences is a fundamental problem with network based packet sequences.

Most of the techniques can operate in an online fashion such techniques not only assigns an anomaly score to a test instance as it arrives. We need to incrementally update the model.

# 6. REFERENCES

[1] Teng, H., Chen, K., and Lu," Adaptive real-time anomaly detection using inductively generated sequential patterns" , Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy. IEEE Computer Society Press, 278-284, 1990.

[2] Rousseeuw, P. J. and Leroy, "Robust regression and outlier detection" , John Wiley & Sons, Inc., New York, NY, USA, 1987.

[3] Tan, P.-N., Steinbach, M., and Kumar, "Introduction to Data Mining", Addison-Wesley, V. 2005.

[4] Phoha, "The Springer Internet Security Dictionary", Springer-Verlag, 2002..

[5] Gwadera, R., Atallah, M. J., and Szpankowski," Detection of signi¯cant sets of episodes in event sequences", Proceedings of the Fourth IEEE International Conference on Data Mining. IEEE Computer Society, Washington, DC, USA, 3-10, 2004.

[6] Gwadera, R., Atallah, M. J., and Szpankowski, "Markov models for identification of significant episodes", Proceedings of 5th SIAM International Conference on Data Mining, 2005a.

[7] SchÄolkopf, B., Platt, J. C., Shawe-Taylor, J. C., Smola, A. J., and Williamson, R. C,".Estimating the support of a high-dimensional distribution", Neural Comput, 13- 7, 1443-1471,2001.

[8] Roth, " Outlier detection with one-class kernel ¯Fisher discriminants. In NIPS.",V.2004.

[9] Roth, "Kernel Fisher discriminants for outlier detection", Neural Computation 18, 4, 942-960, V. 2006.

[10] Agrawal, R. and Srikant," Mining sequential patterns", Proceedings of the 11th International Conference on Data Engineering. IEEE Computer Society , Washington, DC, USA, 3-14, 1995.

[11] Mahoney, M. V. and Chan," Learning nonstationary models of normal network traffic for detecting novel attacks", Proceedings of the 8th ACM SIGKDD international onference on Knowledge discovery and data mining. ACM Press, 376-385, 2002.

[12] Mahoney, M. V. and Chan," Learning rules for anomaly detection of hostile net- work traffic", Proceedings of the 3rd IEEE International Conference on Data Mining. IEEE Computer Society, 601, 2003.

[13] Mahoney, M. V., Chan, P. K., and Arshad, " A machine learning approach to anomaly detection", Tech. Rep. CS,200-306,2003.

[14] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques ",2nd Edition , 2006.