Detecting Outlier of Mass Concrete Temperature During Construction Based on Hybrid Method

Xiaoyun Yu[a], Yihong Zhou[a,*], Chunju Zhao[a,**], Huakang Qin[a]

[a] *China Three Gorges University, 8, Daxue Rd, Xiling, Yichang, Hubei, China*

Abstract

Mass concrete temperature control is the most important work to perform during construction, Commonly, we capture temperature data using Distributed Temperature System (DTS), But the capture process caused by power failure or other defeats lead to a wrong observation results. otherwise, temperature data is a sort of time series, this type records. challenges in outlier detection of this type include unknown data distribution, lack of train data set, multiple parameters and fuzziness of outlier. considering this, a hybrid model is developed whose salient feature is a synergistic combination between fuzzy set based and distance based techniques. This method is suitable for long period time series. the hybrid model set a two stage work flow, in the first stage, considering the fuzziness of outlier, we describe outlier using a fuzzy set, in which defined a threshold to judge whether or not the outlier is, considering the main temperature affection aspect. this algorithm proposed in the paper is distribution-free and robust, throughout the outlier detection method, we try to erase all outliers generated and make plotted temperature curve smoother. A set of experiments show than the method can greatly lower false alarm rate and improve detection efficiently especially in long period temperature time series.

Keywords: Temperature, Outlier Detection, Time Series, Unsupervised Learning

1. Introduction

Outlier detection is able to find patterns in data that do not confirm to expected behavior. It has extensive use in a wide variety of applications such as military surveillance for enemy activities, intrusion detection in cyber security, fraud detection for credit cards, insurance or health care and fault detection in safety critical systems. Detecting outlier of mass concrete temperature during construction is a critical work prior to research and engineering. The reason that we can detect outlier from normal data is *Nature does not make jumps* .

Temperature record is a sort of time series, which is observed or measured at many points in time forms a time series, many time series has a regular fixed time interval, that's to say that data points occur at a regular intervals according to some rules, such as every day, every hour, or every 2 months. Many time series are endless and continuous at a long time. the Temperature time series is a endless time series that data observed per hour. Outlier detection in time series provide significant information for numerous applications. Outlier in time series can manifest in terms of the changes in the amplitude of data, or can be shaped in temporal curve in a axis. there are three types of outlier: point outliers, contextual outliers and collective outliers, in this paper, considering the features of mass concrete temperature data, the prior two types is mainly concerned.

In the temperature time series, It's probably has one type or both type even all type of outliers, sometimes, the point outlier detection problem or collective outlier detection problem and be transformed to contextual outlier detection problem by incorporating the context information.

Outlier detection is related to, but distinct from noise

*Yihong Zhou
**Chunju Zhao
Email addresses: yuxiaoyun@ctgu.edu.cn (Xiaoyun Yu), zhyh@ctgu.edu.cn (Yihong Zhou), little@163.com (Chunju Zhao), qinhuakang1991@qq.com (Huakang Qin)

removal and noise accommodation, 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.

The purpose of this article is to give tips for data users about the possibility of outliers detection, and inform them to response to the possibility of outliers and data failures. This approach implicate mainly in the data observation in the period of dam construction, including grouting, poring, maintaining, and during cooling system working process.

The temperature time series contain continuous data points, which has a unknown distribution, the curve plotted variate from different seasons and each different years.

Mass concrete temperature control is a vital important issue in dam construction, the temperature data sets captured by DTS (Distributed Temperature System), through optics fiber buried in concrete block, the temperature captured by DTS mostly affected by the work conditions, on the dam construction working site with crucial environment often cause the capture equipment run in a wrong way, for example, the power failure will stop distributed temperature system working, and not properly machine moving will affect the accuracy of any facilities. Most of current outlier detection methods are based on statistical. many techniques employed for detecting outliers are fundamentally identical but with different names chosen by authors temperature detection. in statistics, an outlier is an observation point that is distant from other observations.[], In outlier detection, one should select an algorithm that is suitable for their data set in terms []. Temperature data inside the huge concrete block is vital important because temperature fluctuations will change the strain inside block, but mostly, sharp temperature change in a very short time dimension (period) will be removed by most outlier detection model based on mean value and probability. to avoid drop the useful data in outliers detection process, we proposal a improved outliers detection program.

when performing least square fitting to a data serial, it's often best to discard outliers before computing the line of best fit. this is particularly true of outliers along the x direction, since these points may greatly influence the result. but in most time series, the lack of data points will cause the discontinuous time line, and this lead a wrong way to the relative approach.

The most important challenge to temperature time series outlier detection includes unknown data distribution, control limit determination, multiple parameters, training data and fuzziness of 'anomaly'. there was a research about self-adaptive statistical process control on outlier detection improve the false alarm rate. and detection rate and false alarm rate are not affected by different K. fuzziness can effectively reduce false alarm rate.[]

Because the mass concrete temperature time series has a unknown distribution or its distribution is not a normal one, we can not use a traditional statistical model to detect outliers but a hybrid one using statistical model and unsupervised learning model.

2. Methodology

2.1. Definition

2.1.1. Outlier and its common types

An outlier is an observation that lies outside the overall pattern of a distribution. Usually, the presence of an outlier indicates some sort of problem. This can be a case which does not fit the model under study, or an error in measurement. Outliers are patterns in data that do not confirm to a well defined notion of normal behavior. Classification of outliers and non outliers is the main aspect of outlier detection. We classify a data set into 2 classes, outlier points and normal ones(mostly we call it corrective data). In this paper, We focus on mass concrete temperature data in huge dam construction, and try to find any outlier within it. For mass concrete temperature observation, it's a long period work and commonly under unstable circumstance, so the capture data is not always fair and stable, large scale data processing leads to man power impossibility, for the benefit of data analysis, we must deal with the raw data and get rid of obvious outlier. The main models in the approach is classification. We do not prefer to arise a program to clean outlier from the data set, instead of remind the data users about the data fault probability. Outlier detection is far more different from novelty detection, noise removal and other techniques dealing with abnormal input data. Noise removal is dealing 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; Novelty aims at detecting previously unobserved (emergent, novel) patterns in the data; but outlier detection is detecting outliers which are patterns that deviate from expected normal behavior, which in its simplest form could be represented by a region and visualize all normal observations to belong to this normal region and consider the rest as outliers. Typically, outliers detection can be classified into three categories: point outliers detection, contextual outliers detection and collective outliers detection []. Outlier can also occur when comparing relationships between two set of data. Outlier of this type can be easily identified on a scatter diagram.

2.1.2. The cause of outlier

- Note taken on *[2016-09-01 Thu 15:01]*
  statistics in outlier cause. the percentage comes from a sample of one year's period.

Generally speaking, unstable observe method, not properly observe job and abominable environment affect the observation, this leads to outlier. We use DTS to observe the temperature inner mass concrete block, DTS fault tolerance performance is not as good as expected, and the quality of fiber burying technique will affect the performance of DTS, In mass concrete temperature capturing, sometimes the power failure and mistake in operation can cause lost in a time series. There are 5 main cause of outlier, see table 1.

2.1.3. Definite Outliers in time series

Outliers in time series that are the least similar to in all other data sets. Commonly, we define the following elements of a time series:

1. Time stamps, which specify a instants in time.
2. Fixed periods, such as the month March 2015 or the full year 2016.
3. Intervals of time, indicated by a start and end time stamp. Periods can be thought of as special cases of intervals.

2.1.4. Characteristic of my research objects

The mass concrete temperature time series is a directive sequence, see figure Through **model character**:

1. Non-parametric model. co-variance or spectrum the process without assuming that the process has any particular structure.
2. Unknown distribution model. unsupervised method and semi supervised model.
3. Time series model. Classify and label normal data or anomalous value.
4. Sequences data, a continuous and endless series.
5. Change rate and maximum value are the most important aspects. *Nature does not make a jump.*
6. Bayes equation. statistical model.

2.2. Current outlying detection methods

Each data instance can be described using a set of attributes, the attributes can be different types such as binary, categorical or continuous. each data instance might consist of only one attribute or multiple attributes.

2.2.1. statistical method (average standard variation)

Statistical approaches were the earliest algorithms used for outliers detection [,]. There are proximity-based techniques, parametric methods, non parametric methods and semi-parametric methods. For liner data outliers detection, commonly detect process is:

1. Separate the data series into several fixed length subsequences;
2. For every sub-sequence, calculate the mean, standard square deviation;
3. Judge the point by calculate the difference between ϵ and ξ to find the outliers.

The judgment standard is the formulation 1

$$\sum_{i=1}^{n}(y_i - y_0)^2 \tag{1}$$

Time series that stay in a state of statistical process control are called in-control data (normal data), otherwise, are called out-of-control data (outliers), we use a control chart to determine if a process is in a state of statistical control. as shown in figure 4.

One of the main limitations of distribution based approaches is: information about the underlying data distribution may not always be available.

Table 1: Mass concrete temperature outlier causes and outlier types

ID	Cause	Outlier Type	Percentage(%)
1	Power failure or voltage fluctuation	Discontinuous curve	30
2	Facility failure or not properly working	Point outlier	20
3	Instrument maintaining	Discontinuous curve, Collective outliers	10
4	Dust coverage in the fiber joint poor contact	Contextual Outliers, collective outliers	10
5	Other errors and other defeats	Point outlier, contextual outlier, collective outliers	10

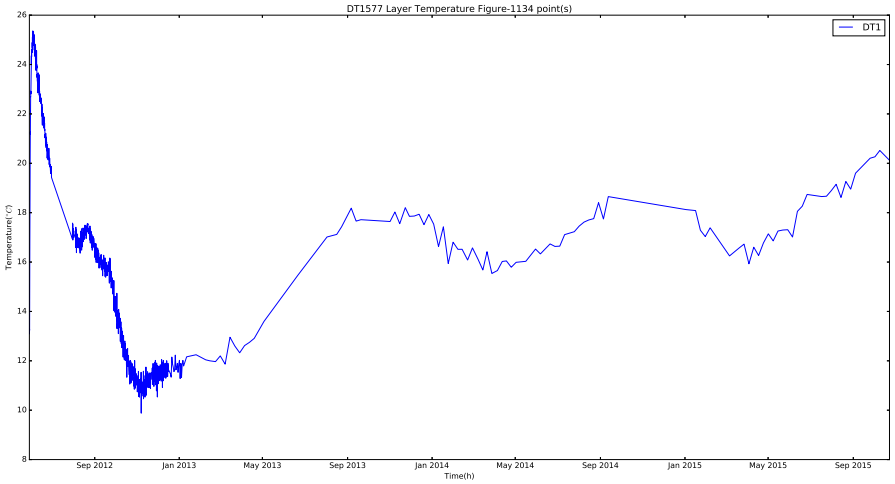

Figure 1: Mass concrete temperature series

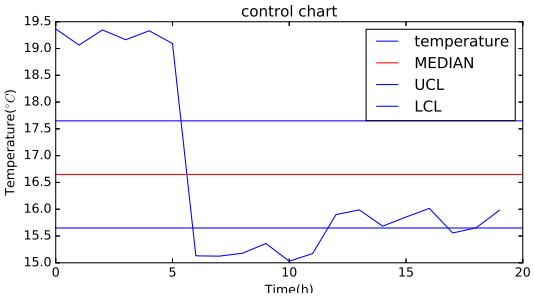

Figure 2: Control Chart

2.2.2. Distance and/or density based methods

This method assume that objects may be outlier relative to their local neighborhood. This triggered interest in development of many variants of the algorithms, which are more spatially oriented.

The famous univariate case, the z-score is a distance-based measure that can be defined as a standard residual:

$$z_i = \frac{|p_i - \overline{p}|}{StD(p)}, i = 1, \ldots, k \tag{2}$$

Where $\overline{p}$ and $StD(p)$ are the mean and standard deviations (StD) of the variable P, In the univariate model, the mean and StD are the median and Median Absolute Deviation (MAD), which is defined as the median of the absolute deviations from the data's median, that's $MAD = median(|X_i - median(X)|)$, so we transform the equation 2 into the following:

$$Mz_i = \frac{|p_i - median_j(p_j)|}{MAD(p)}, i = 1, \ldots, k \tag{3}$$

$median_j(p_j)$ is a more robust mean as $\overline{p}$, $MAD(p)$ is standard deviation as $StD(p)$ in the equation 2. Robust approaches have been developed to avoid/reduce the outlier influence on the estimates.[]

2.2.3. Clustering based method

Clustering based methods apply unsupervised clustering techniques mainly to group the data based on their local data behavior.

2.2.4. Model based approach

Model based approach that is used to learn a model(classifier) from a set of known data, i.e. training data, and then categories the data into inliers and outliers. This method can detect outliers in a high-dimensional data but require much more time to construct a classifier.[]

2.2.5. Neural networks

there are several tries in outliers detection using neural networks, Simon Hawkins using replicator neural networks to accomplish outliers detection.[] supervised neural methods and unsupervised neural methods are used in outliers detection.

in this study, we try to pursuit a way to verify the data observed is reliable, or if we are sure the data is suitable after the outlier detection process.

for time series, the data last for a continuous series, if any point observation has a fault value, or lack of points because of detection equipment power failure. discontinued temperature time series caused by incomplete captured data lead dam concrete block temperature adjust and control.

2.2.6. Machine learning method

Most of outlier detection has been performed exist in statistics. However, many outlier detection approaches have been developed in machine learning, pattern recognition and data mining and are referred to by different names e.g. novelty detection, anomaly detection, exception mining or one-class classification. Much outlier detection has only focus on continuous real-valued data attributes there has been little focus on categorical data. most statistical and neural approaches require cardinal or at least ordinal data to allow vector distances to be calculated and have no mechanism for processing categorical data with no implicit ordering. Skalak and Rissland use a C4.5 decision tree to detect outliers in categorical data and thus identify errors and unexpected entries in data tables.[] decision trees do not require any prior knowledge of the data unlike many statistical methods or neural methods that require parameters or distribution models derived from the data set. One must be sure that input data attributes or features have sufficient information to predict the outputs. and after model setup, one would like to get insight by identifying attributes which are influencing the output in a significant manner.

2.2.7. Hybrid Systems

The most recent development in outlier detection technology is hybrid system. These systems incorporate algorithms from at least two of the preceding detection methods. Hybridization is used variously to overcome shortage of efficiency with one particular classification algorithm, to exploit the advantages of multiple approaches while overcoming their weaknesses or using a meta-classifier to reconcile the outputs from multiple classifiers to handle all situations. We describe approaches where an additional algorithm is incorporated to overcome weaknesses with primary algorithm next.

2.3. Introduction of the method we proposed

In the paper, We propose a hybrid resolvement combine statistical algorithm and a distance based computation algorithm. statistical method is more accurate in experimental data, it's a traditional method to detect continuous time series. Distance based computation detect the burst sharp change points that does not follow statistical normal distribution. The structure of this method depicted in figure 3.

We defined labels for points in a data set. The labels associated with a data instance denote if that instance is normal or outlier. It should be noted that obtaining labeled data which is accurate as well as representative of all types of behaviors, is often prohibitively expensive. Labeling is often done by a human expert and hence requires substantial effort to obtain the labeled training data set, typically, getting a labeled set of outlier data instances which cover all possible type of outlier behaviors is more difficult than getting labels for anomalous one. The most important thing is that in mass concrete temperature data set, we can label neither normal type nor anomalous data.

All things we focus on is the fluctuations rate and maximum value (This commonly call double control) in mass

concrete temperature. We detect outliers for maximum values using a traditional statistical model, and a distance based method to detect fluctuation rate outliers.

The technique details:

1. Compute outlier scores;
2. Label outlier and normal.

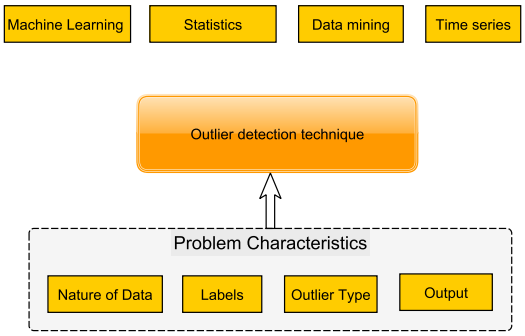

Figure 3: Outlier detection schematic diagram

The most difficult thing to detect outlier is defining the normal data or normal region, and what's more, different notion of outliers in different application domains makes it difficult to apply techniques developed in one domain to another. Besides, there is few labeled data for training in mass concrete temperature outlier detection.

2.3.1. Procedure for detecting outlying observation in samples

Once some of the sample observations are branded as 'outliers', then a thorough investigation should be initiated to determine the cause. In particular, one should look for gross errors, personal errors, errors of measurement, errors in calibration, etc. If reasons are found for aberrant observations, then one should act accordingly and perhaps scrutinize also the other observations. Finally, if one reaches the point that some observations are to be discarded or treated in a special manner based solely on statistical judgment, then it must be decided what action should be taken in the further analysis of the data. []

2.4. A control chart of rudimentary detection

Statistical process control (SPC) is a method of quality control which uses statistical methods. SPC is applied in order to monitor and control a process. Monitoring and controlling the process ensures that it operates at its full potential. At its full potential, the process can make as much conforming product as possible with a minimum (if not an elimination) of waste (rework or scrap). SPC can be applied to any process where the "conforming product" (product meeting specifications) output can be measured. Key tools used in SPC include control charts; a focus on continuous improvement; and the design of experiments. An example of a process where SPC is applied is manufacturing lines.[] As it's said before, a control chart model is easy to apply in a time series, in this approach, we use a control chart to narrow the variation of the data set. The following figure show control chart works on the temperature time series data set. figure 4 show how to limit the data into a normal zone, and if the points located out of the range between UCL and LCL, It's probably a outlier.

Let us consider a time series X, which is an sequence taken from mass concrete temperature data set, n is the number of sub-sequence in X, P is one of X with a collection of q features. The key factor of an outlier detection algorithm is the similarity used to determine how closely two given sub-sequences are matched. Here, we measure similarity between two Ps using Euclidean distance.

For each sub-sequence, we define the degree to which the two sub-sequence defer to each other as outlier score. the outlier score is described as the weighted average of the Euclidean distance between one sub-sequence and its K nearest neighbors, formally speaking, the outlier score represented as the following formulation:

$$Outlierscore(P) = \frac{1}{K}\sum_{t=1}^{K} Distance(P, P_t)|P_t \in KNNSet(P) \tag{4}$$

where $KNNSet(P)$ is the set of K nearest neighbors of P in time sequence series, $Distance(P, P_t)$ is Euclidean distance.

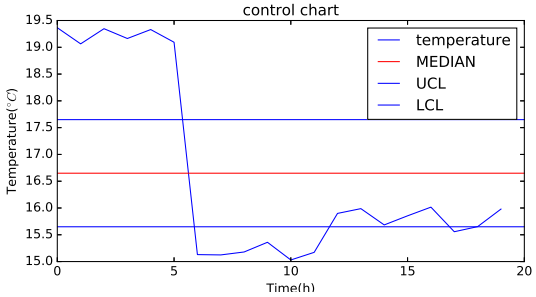

Figure 4: A Control Chart Figure

2.4.1. Distance based method

A more robust and flexible technique considers a set of k nearest neighbors instead of one, not surprisingly it is called k-nearest-neighbors (KNN). The flexibility is given by different possible classification techniques. For example, the output can be that of the majority of the k neighbors outputs. If one wants to be on the safer side, one may decide to classify the new case only if all k outputs agree (unanimity rule), and to report "unknown" in the other cases the data can be created by the previous history of the optimization process or by feedback by the decision makers.[,]

The distance based method require a user-supplied parameter k to fix the size of the neighborhood set (which will henceforth be referred to as the 'kNN set' or 'kNNs, and the distance to the kth nearest neighbor will be referred to as 'k NN distance'). Since finding the 'kNNs' in a data set of size N = |DB| by means of a linear scan requires $O(N^2)$ distance computations (which can often be reduced to $O(NlogN)$ using appropriate index structures), the cost of finding the neighbors usually dominates the algorithm run-time []. through a experimental evaluation, Campos et al. [] suggest that novel methods should not be proposed without also indicating those domains or application scenarios where this method is particularly well suited. Merely demonstrating that the method excels on a few data sets for a few parameter settings does not suffice. Most importantly, broad ranges of parameter choices should be tested for the competitors.

The value at a unknown point should be the average of the known values at its neighbors, weighted by the neighbors' distance to the unknown point. That is to say, some information about the function to be optimized is missing at the beginning, and only the decision maker will be able to fine-tune the search process, solving many if not most real world problems requires interactive processes with learning involved [].

The user will learn and adjust his preferences after knowing more and more cases, the system will build models of the user preferences from his feedback. the steps will continue until the user is satisfied or the time allocated for the decision is finished.[]

Natura non facit saltus means that Nature does not make jumps. nature things and properties change gradually, rather than suddenly. this is the basic of outlier detection technology.

The main purpose of outlier detection is if captured data points report 'unknown', the safer response to the unknown point is that can be suggested to remind the user's attention.

Let $k \leq l$ be a fixed positive integer (l is the number of labeled examples), and consider a feature vector x. a simple algorithm to estimate its corresponding outcome y consists of two steps:

1. find within the training set the k indices $i_1, \ldots, i_k$ whose feature vector $x_{i_1}, \ldots, x_{i_k}$ are nearest to the given x vector.
2. Calculate the estimated outcome y by the following average, weighted with the inverse of the distance between feature vector:

$$y = \frac{\sum_{j=1}^{k} \frac{y_{i_j}}{d(x_{i_j},x)+d_0}}{\sum_{j=1}^{k} \frac{1}{d(x_{i_j}),x} + d_0} \tag{5}$$

Where $d(x_i, x)$ is the distance between the two vectors in the feature space (for example the Euclidean distance) and d_0 is a small constant offset used to avoid division by zero. The larger d_0, the larger the relative contribution of far away points to the estimated output. if d_0 goes **infinity**, the predicted output tends to the *mean* output over all training examples.

Real learning is associated with extracting the deep and basic relationships in a phenomenon, with summarizing with short models a wide range of events, with **unifying different cases by discovering the underlying explanatory laws**.

feature extraction from original data observed by tools into featured input x

1. the first step is getting individual measurable properties of the phenomena being observed with useful information to derive the output value.
2. the second step is obtain the parameters between x and y using automated machine learning method.

Identifying the best model requires identifying the proper 'model complexity'.

Bayes' theorem: **conditional** probability density $p(x|y)$

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)} re \tag{6}$$

2.4.2. Outlier detection as statistical hypothesis testing

Traditional hypothesis testing relies more on prior knowledge of the data distribution, so these type of test are not suitable for arbitrary or new and unknown data sets. What's more hypothesis testing is a statistical estimation based on distribution features, because we do not know distribution of the temperature time series data, that's to say, we can not estimate its parameters, but each point distributed in the time series is not coincide with the characteristic of temperature properties. In our hypothesis model, we consider to set up a threshold to do hypothesis testing to determine whether one point is a outlier or not. We judge by identifying the state created by comparing anomaly score and the threshold set prior. We assume the hypothesis H_0 as the outlier, and H_1 as the normal data. formula as equation 7.

$$\begin{aligned} H_0 &: score > threshold \\ H_1 &: score \leq threshold \end{aligned} \tag{7}$$

2.5. Supervised learning in different temperature phase

We separate the full time series into different part base on the different stage and environment. There are 3 main features that affect the time series values, the first is temperature control stage. We separate the whole temperature time series into:

1. Temperature rising stage.
2. Temperature control stage.
3. Cooling stage.
4. Suspend stage.
5. Temperature stable stage.

Please refer to figure 5

Through the observed history data, we can separate them identified by the characteristic point. the features are temperature value ($°C$) and temperature variation rate ($°C/h$)

Figure 5: Temperature change and control stage

2.6. Efficiency and Effectiveness

Researchers are paying much interest in outlier detection efficiency and effectiveness. With regard to efficiency, the distance based outlier detection models rely heavily on many different factors, such as data set size and dimensionality, the choice of parameters, the data structures organized, and other implementation features. in this paper, we evaluate the efficiency on the method proposed. through a fixed data structure but a fluctuate parameters to choose the better efficiency. On the other hand, the effectiveness of algorithms for outlier detection has less focus than efficiency issues. In this paper, we discuss the effectiveness through different data samples and different type of parameters [].

2.7. Summarize

2.7.1. Difficulties in outlier detection

1. It's very difficult to define the normal behavior or a normal region.
2. Imprecise boundary between normal and outlier behavior since at times outlier observation lying close to the boundary could actually be normal, and vice-versa.

2.7.2. Paper structure

The paper is organized as the following. Section 2 illustrates a fuzzy based model connected to statistical method,

in section 3, we come out a Control chart approach to limit the outlier scope. section 4, we combine a distance based model and a statistical model to detailed those temperature time series outlier detection. Finally, we made conclusion and sumarize to the whole paper in section 5$.

3. Experiments and Discussions

We begin experiments by showing the usefulness of the proposed algorithm for synthetic data and real life data, including outliers in shape and amplitude. We contrast our algorithm against several baseline algorithms to show that the proposed algorithm is able to efficiently find outliers. In our experiments, it is workable without training process. K can assume any value. The hybrid outlier detection method between probability and distance based that we proposed in this article can largely find temperature anomaly data series.

3.1. Effectiveness

Give a example, when detect through these process, and find if the outliers can be detect by our model.

1. use a artificial data set which add any anomaly points to test.
2. Use a real world temperature time series to test efficiency.

3.1.1. Curves in ROC (Receiver operating characteristic) spaces

In binary classification, the class prediction for each instance is often made based on a continuous random variable X, which is a "score" computed for the instance (e.g. estimated probability in logistic regression). Given a threshold parameter T, the instance is classified as "positive" if $X > T$, and "negative" otherwise. X follows a probability density $f_1(x)$, if the instance actually belongs to class "positive", and $f_0(x)$ if otherwise. Therefore, the true positive rate is given by $TPR(T) = \int_T^\infty f_1(x)dx$ and the false positive rate is given by $FPR(T) = \int_T^\infty f_0(x)dx$. The ROC curve plots parametrically $TPR(T)$ versus $FPR(T)$ with T as the varying parameter. $TPR(T) = \int_T^\infty f_1(x)dx$

3.2. Performance analysis

We now analysis the performance of some techniques in our algorithm. A key feature of our algorithm is a combination of both statistical-based techniques and distance-based algorithm.

3.3. Comparison with other methods

For cumulative effectiveness, distribution function is unknown, statistical nature is lying on the surface of the data.

4. Conclusion

We proposal a DTS data sets outlier detection and retention method using a combination of statistical and 3 alpha. It's more accurate and speed up. the algorithm is robust when dealing with large data flow. learning from examples is only a means to reach the real goal. Thought the work can largely identified most of outliers in large concrete dam block temperature time series, but it is a distance-based method, can not apply in other time series. The performance may be significantly affected by different proportion of outlier data in data sets. The future work will focus on the nature of mass concrete temperature change and discovery of concrete construction work arrangement.

5. Acknowledgments

The work of this work is supported by the project of National Natural Science Foundation of China (No.51379109 and No.51479103).

6. References