Rare Pattern Detection By Using Weather Forecasting Framework

Authors
Shraddha V. Shingne¹, Prof. Anil D. Warbhe², Prof. Shyam Dubey³

¹M.Tech II year Dept. of Computer Science and engineering, Nuva College of engineering and technology, Nagpur, Maharashtra,
²Asst. Prof Dept. of Electronics engineering, Manoharbai Patel institute of engg. and technology, Gondia, Maharashtra
³H.O.D. Dept. of Computer Science and engineering, Nuva College of engineering and technology, Nagpur, Maharashtra

Email- s.shingne90@gmail.com, anilwarbhe@hotmail.com, Shyam.nuva@rediffmail.com

ABSTRACT
India has a typical weather conditions consisting of various seasons and geographical conditions therefore we need to study weather condition. It is necessary to predict climate change and meteorology observes the changes in temperature, air pressure, wind direction and moisture. For weather forecast it is necessary to first analyze and simplify the data before proceeding with other analysis. To detect the outlier we used the adaptive clustering and K-means algorithm. The goal of the proposed method is to implement a powerful, efficient and adaptive rare pattern detection technique consisting of four parts missing data recognition, data preprocessing, initial clustering and adaptive clustering for dynamic detection. This method is implemented using mat lab which uses metrological data in the form of daily data, collected from government data center, taken as input in Microsoft excel sheet. The result is generated in the form of histogram. Different weather patterns are reveal in the result, existing depending on the analysis of historical metrological data. With the help of these histograms we can easily identify the occurrence of rare patterns in weather which helps us to predict future environmental changes in the climate.

Keywords: Data Mining, Data Mining Techniques, weather data, meteorological data.

I. INTRODUCTION

In this paper, we introduce techniques to understand and detect the patterns of climate change by using data from daily weather records. Pattern detection, especially anomaly pattern detection, as a data mining task, refers to disclosing patterns that do not conform to expected behaviours in databases. Climate change is a widely recognized global environmental challenge [1]. These unusual patterns are often referred to as
different terms in different application domains, such as rare patterns [2], outliers [3], [4], faults [5], peculiarities or contaminants, and etc [6].

Most existing pattern detection techniques resolve specific formulations of a problem. The formulations are induced by various factors such as the nature of the data, availability of the labeled data, and the type of anomalies to be detected, and etc. A considerable amount of data mining research on pattern detection has been conducted, and this stream has gained considerable interest owing to the realization that anomaly patterns can be detected from very large databases by data mining [4]. In addition, pattern detection is normally carried out through fault detection methods which analyze the current signals obtained from induction machines [5]. High-dimensional current signals are transformed to low-dimensional data by the mapping of the original signals into different clusters according to their characteristics [5].

Pattern detection methods in the field of data mining typically pick out different patterns (clusters), their changes, and rare/outlying events and such changes are often the source of problems in impact studies [6]. They have been successfully applied to many fields. However, only a few studies have been adapted for environment, weather or climate applications.

From a meteorological viewpoint, research on climate change to disclose typical seasonal patterns of weather, such as the seasonal variability of thermal conditions in Singapore is very important [7], [8]. Natural variability in the climate of Singapore is influenced primarily by the monsoonal influences in various months of the year. Studies on Singapore data have shown a long-term increase in temperature in the past few decades [9], [10], in line with global warming studies over the Southeast Asian region [11]. The anomaly pattern or rare events may also be attributed to anomalous conditions such as the El Nino or Southern Oscillation (ENSO) phenomena [12]–[13].

On a larger scale, climate change is an unprecedented environment change that is affecting our planet [14], [15]. It is already having significant impacts on many aspects of our lives [14], [16]–[19]. Climate change projections on both the global and regional scales are characterized by multiple sources of uncertainty [16]. In order to characterize such uncertainties, global and regional climate model projections need to be based on probabilistic approaches using multimodel ensembles of experiments [16].

Kyung Soo Jun et al. discussed the impact of climate change on spatial water resources. The study was a new attempt to quantify hydrologic vulnerability that included the impacts of climate change [17]. The long term impact of climate change on the carbon budget of Lake Simcoe, Ontario were discussed by analyzing the relationship between temperature and dissolved inorganic carbon in some tributaries [19]. Anna Augustsson et al. discussed how the climate effect can be inserted in a commonly used exposure model, and how the exposure then changes compared to present conditions [18]. The results indicated that changes in climate are likely to affect the speciation, mobility, and risks associated with metals.
As mentioned previously, many research activities have been carried out to improve the understanding of climate change patterns by means of different techniques and made considerable contributions [19]–[20]. However, the introduction of data mining techniques into this research field has been limited. Therefore, this paper aims to develop a detection method based on data mining techniques for detecting and classifying weather patterns through a case study on weather data.

In this paper, we propose an adaptive clustering pattern detection data mining method based on an incremental cluster chain. The proposed method, an adaptive clustering pattern detection method, has a flexible structure for allowing the pattern grows and the clusters are adjusted adaptively. The proposed method is applied to analyze the weather patterns through meteorological data mining and different weather patterns are disclosed. The results indicate that the early weather patterns disappear consistently across models and this suggests long-term climate changes. The proposed pattern detection algorithm will be of potential use for climatic and meteorological research as well as research focusing on pattern recognizes or knowledge discovery in other research field.

II. OBJECTIVE

The objective of the proposed method is to detect the climate change patterns through meteorological data mining. Meteorological variables, including longitude, latitude, mer wind, zone wind, humidity, air temperature and sea surface temperature are simultaneously considered for identifying climate change patterns. Different scenarios with varied cluster thresholds are employed for testing the sensitivity of the proposed method. The robustness of the proposed method is demonstrated by the results. It is observed from the results that the early weather patterns that were present in past disappear consistently across models. Changes in temporal weather patterns suggest long-term changes to the global climate which may be attributed in part to urban development, and global climate change on a larger scale. Our climate change pattern detection algorithm is proven to be of potential use for climatic and meteorological research as well as research focusing on temporal trends in weather and the consequent changes. The research promote the cause of advanced study and research and to effect coordination of research and investigations in all disciplines related to weather forecasting.

Ultimately the Objective is to:
“Detect weather patterns in the long-term while consistent with global climate change on weather patterns adaptively[21]”.

III. PERPOSED METHODS

Collection of Data

Meteorological data in the form of daily summary data were extracted from National Climate Data
Knowledge of meteorological data in a site is essential for meteorological, pollution and energy applications studies and development. Especially temperature data is used to determine thermal behavior (thermal and cooling loads, heat losses and gains) of buildings. It is also an explicit requirement for sizing studies of thermal and/or PV systems. Another major sector where temperature data is fundamental is the estimation of biometeorological parameters in a site. In advanced energy system designs the profile of any meteorological parameter is a prerequisite for systems operating management on daily and/or hourly basis. Also, simulations of long-term performance of energy plants require detailed and accurate meteorological data as input. This knowledge may be obtained, either by the elaboration of data banks, or by the use of estimation methodologies and techniques, where no detailed data are available. As nowadays “smart buildings” have become a reality, artificial techniques must be embedded in building management systems (BMS), in order energy profile (loads, gains etc) of a following time period (next hour, next day) to be predetermined. That will lead to a more effective energy management of the building or the energy plant. Weather data from automated weather stations have also become an important component for prediction and decision making in agriculture and forestry. The data collected from such stations are used in predictions of insect and disease damage in crops, orchards, turfgrasses, forests and weather patterns; in deciding on crop-management actions such as irrigation, in estimating the probability of occurrence of forest fires, and in many other applications. Errors in weather data are frequently caused by poorly calibrated instrumentation, but also may be caused by errors in recording the data or while digitising older hard-copy records. In either case, this data must be “cleaned” before accurate valuation analyses can be performed on Weather data. Meteorological data in the form of daily summary data were extracted from National Climate Data Center(NCDC), National Oceanic and Atmospheric Administration(NOAA).

Table 1: Metrological data in excel sheet

<table>
<thead>
<tr>
<th>Obs</th>
<th>Yr</th>
<th>Mont h</th>
<th>Day</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Zone. Winds</th>
<th>mer. winds</th>
<th>Humidity</th>
<th>Air temp.</th>
<th>s.s. Temp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>80</td>
<td>3</td>
<td>7</td>
<td>.</td>
<td>-139.96</td>
<td>-5.9</td>
<td>-3.4</td>
<td>88.9</td>
<td>27.75</td>
<td>28.71</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>3</td>
<td>8</td>
<td>-2.03</td>
<td>-139.97</td>
<td>-4.5</td>
<td>-5.2</td>
<td>87.7</td>
<td>27.08</td>
<td>28.27</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>3</td>
<td>9</td>
<td>-2.02</td>
<td>-139.97</td>
<td>-3.6</td>
<td>-4.5</td>
<td>83.1</td>
<td>26.6</td>
<td>27.7</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
<td>3</td>
<td>10</td>
<td>-2.02</td>
<td>-139.96</td>
<td>-2.6</td>
<td>-2.3</td>
<td>84.4</td>
<td>26.41</td>
<td>27.27</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>3</td>
<td>11</td>
<td>-2.02</td>
<td>-139.96</td>
<td>-1.2</td>
<td>-5.1</td>
<td>88.5</td>
<td>25.93</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>3</td>
<td>12</td>
<td>-2.01</td>
<td>-139.96</td>
<td>-4</td>
<td>-4.3</td>
<td>84.4</td>
<td>25.54</td>
<td>26.77</td>
</tr>
<tr>
<td>7</td>
<td>80</td>
<td>3</td>
<td>13</td>
<td>-2.01</td>
<td>-139.96</td>
<td>-4</td>
<td>-3.1</td>
<td>84.8</td>
<td>25.69</td>
<td>26.87</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>3</td>
<td>14</td>
<td>-2.02</td>
<td>-139.96</td>
<td>-2.8</td>
<td>-4.5</td>
<td>89.3</td>
<td>25.61</td>
<td>26.86</td>
</tr>
<tr>
<td>9</td>
<td>80</td>
<td>3</td>
<td>15</td>
<td>-2.02</td>
<td>-139.97</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>25.26</td>
<td>26.55</td>
</tr>
</tbody>
</table>

Shraddha V. Shingne, et al IJSRE Volume 3 Issue 7 July 2015
Missing Values Recognition

A very common difficulty is that the value of a given feature is missing for some training and/or test examples. Often, missing values are indicated by question marks. However, often missing values are indicated by strings that look like valid known values, such as 0 (zero). It is important not to treat a value that means missing inadvertently as a valid regular value. Some training algorithms can handle missingness internally. If not, the simplest approach is just to discard all examples (rows) with missing values. Keeping only examples without any missing values is called “complete case analysis.” An equally simple, but different, approach is to discard all features (columns) with missing values. However, in many applications, both these approaches eliminate too much useful training data. Also, the fact that a particular feature is missing may itself be a useful predictor. Therefore, it is often beneficial to create an additional binary feature that is 0 for missing and 1 for present.

While a poorly constructed series can contain empty values, it is very unlikely. Missing values cause a major problem in series data, and it is very unlikely that any series would be constructed to permit empty values. In any case, whether missing or empty, series modeling techniques fare even worse with values that are absent than non-series techniques. There are two dimensions of a series in which a value could be missing: the feature variable, and the index variable. Here attention will be confined to a value missing in the feature variable. When replacing a missing value in non-series data, joint variability is preserved between variables, and a suitable value for the replacement is found using the information that is contained in whatever variable values are present.

Incomplete data is an unavoidable problem in dealing with most of the real world data sources. This has been discussed and analyzed by several researchers. Depending on the case, the expert has to choose from a number of methods for handling missing data.

**Method of Ignoring Instances with Unknown Feature Values:**
This method is the simplest, just ignore the instances, which have at least one unknown feature value.

**Most Common Feature Value:**
The value of the feature that occurs most often is selected to be the value for all the unknown values of the feature.

**Concept Most Common Feature Value:**
This time the value of the feature, which occurs the most common within the same class is selected to be the value for all the unknown values of the feature.

**Mean substitution:**
Substitute a feature’s mean value computed from available cases to fill in missing data values on the remaining cases. A smarter solution than using the “general” feature mean is to use the feature mean for all samples belonging to the same class to fill in the missing values.

If a feature with missing values is retained, then it is reasonable to replace each missing value by the mean or mode of the non-missing values. This process is called imputation. More sophisticated imputation procedures exist and they are mostly better.

There can be multiple types of missingness, and multiple codes indicating a missing value. The code NA often means “not applicable” whereas NK means “not known.” Whether or not the value of one feature is missing may depend on the value of another feature. For example, if wind speed is zero, then wind direction is undefined.

C. K-means

K-means clustering is a partitioning based clustering technique of classifying/grouping items into k groups (where k is user specified number of clusters). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid. A centroid (also called mean vector) is "the center of mass of a geometric object of uniform density". Although K-means is simple and can be used for a wide variety of data types, it is quite sensitive to initial positions of cluster centers. There are two simple approaches to cluster center initialization i.e. either to select the initial values randomly, or to choose the first k samples of the data points. As an alternative, different sets of initial values are chosen (out of the data points) and the set, which is closest to optimal, is chosen. However, testing different initial sets is considered impracticable criteria, especially for large number of clusters, Is mail et al (1989). Therefore, different methods have been proposed in literature by Pena et al. (1999). Also, the computational complexity of original K-means algorithm is very high, especially for large data sets. Computer science has been widely adopted in different fields like agriculture. One reason is that an enormous amount of data has to be gathered and analyzed which is very hard or even impossible without making use of computer systems. The research of spatial data is in its infancy stage and there is a need for an accurate method for rule mining. Association rule mining searches for interesting relationships among items in a given data set. This method enables us to extract pattern.

K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. It classifies a given set of n data objects in k clusters, where k is the number of desired clusters and it is required in advance. A centroid is defined for each cluster. All the data objects are placed in a cluster having centroid nearest (or most similar) to that data object. After processing all data objects, k-means, or centroids, are recalculated, and the entire process is repeated. All data objects are bound to the clusters based on the new centroids. In each iteration centroids change their location step by step. In other
words, centroids move in each iteration. This process is continued until no any centroid move. As a result, k clusters are found representing a set of n data objects.

**K-means Algorithm:**

1) Specify k, the number of clusters to be generated
2) Choose k, points at random as cluster centers
3) Assign each instance to its closest cluster center using Euclidean distance
4) Calculate the centroid (mean) for each cluster; use it as a new cluster center
5) Reassign all instances to the closest cluster center
6) Iterate until the cluster centers don’t change anymore.

For each of the initial clusters specified, the random centroid of the $i^{th}$ cluster with m−1 number of input data is represented as following:

$$C_{i}^{(m-1)} = \sum_{j=1}^{m-1} \frac{x_j}{m-1}$$  \hspace{1cm} (1)

The similarity distances for the numeric values of attributes and similarity count for the number of attributes is set commonly for each initially formed clusters and if the new instance of data satisfy both conditions then each new instance of the data is assigned to its closest cluster center using Euclidean distance formula. The Euclidean distance is employed to calculate the distance between the two input data vectors:

$$X_m = (X_m(1), X_m(2), \ldots, X_m(n)) \text{ and } C_{i}^{(m-1)} = (C_{i}^{(m-1)}(1), C_{i}^{(m-1)}(2), \ldots, C_{i}^{(m-1)}(n))$$ \hspace{1cm} (2)

in n-dimensional space:

$$D(X_m, C^{(i)}) = \sqrt{\sum_{j=1}^{m-1}(X_m(j) - C^{(i)}(j))^2}$$ \hspace{1cm} (3)

The centroid of the new formed cluster for each new instance of data is updated every time using adaptive clustering:
D. Adaptive Cluster Formation

In the proposed method as the initial cluster formation is done on the training data selected using k means algorithm and Euclidean distance, further each instance of the testing data is selected and each numeric value of the attribute compared with the distance calculated with the centroids of the initial clusters. The similarity distance threshold is set for attributes values as well as attributes count for each initially formed cluster again. If the new instance satisfies the distance thresholds, it will be added to one of the initial clusters according to the distance with centroids. If not, the new cluster is formed. This process continues till the end of testing data available.

Adaptive Clustering Method:

1. Specify No of new clusters other than initial clusters.
2. Choose each new instance from the testing data as an input
3. Set the different similarity threshold for distance with the centroids and attributes count.
4. Compare the new instance with existing clusters according to the distance similarity threshold.
5. If satisfies the similarity with the centroids of one of the initial clusters, add to that particular cluster.
6. If doesn’t satisfies the similarity, a new pattern exists
7. till the end of data.

IV. RESULTS AND DISCUSSIONS

The data is made train and apply preprocessing technique to remove the missing values. The preprocessed data is loaded successfully and it is shown in the following module (Fig.1). Now number of many missing values present in data are evaluated and generated in report format module (Fig.2).

![Figure 1: Data Model](image-url)
Figure 2: Missing Value report module

Figure 3 shows the result in the form of histogram. After the application of adaptive clustering method, some new clusters are formed whose data doesn't belong to the initially created clusters according to Euclidean distance applied. Each bar in the histogram represents a pattern and the bars having lowest density shows the rare patterns detected.

Figure 3: Histogram of 9 weather patterns

Figure 4: All the clusters with values.

Figure 4 shows all formed clusters after the application of adaptive clustering technique and values belonging to initial as well as later formed clusters.
V. CONCLUSION

The proposed method is capable of simultaneously dealing with all meteorological or climate variables to detect hidden weather patterns. By the result of proposed algorithm an Adaptive K-means Algorithm is able to detect patterns and their rare outlying changes. The ability to forecast the weather accurately is an increasingly important part of our economy and our society. This method presents an efficient data processing technique used to enable the available data proper for the knowledge discovery as if the available metrological data is not proper the detection can be improper which leads to improper prediction of climate change.

REFERENCES


