Multivariate weather anomaly detection using DBSCAN clustering algorithm

Multivariate weather anomaly detection using DBSCAN clustering algorithm

S Wibisono*, M T Anwar, A Supriyanto and I H A Amin

Faculty of Information Technology, Universitas Stikubank, Jl. Tri Lomba Juang No 1 Semarang 50241, Central Java, Indonesia

*setyawan@edu.unisbank.ac.id

Abstract. Weather is highly influential for human life. Weather anomalies describe conditions that are out of the ordinary and need special attention because they can affect various aspects of human life both socially and economically and also can cause natural disasters. Anomaly detection aims to get rid of unwanted data (noise, erroneous data, or unwanted data) or to study the anomaly phenomenon itself (unusual but interesting). In the absence of an anomaly-labeled dataset, an unsupervised Machine Learning approach can be utilized to detect or label the anomalous data. This research uses the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to separate between normal and anomalous weather data by considering multiple weather variables. Then, PCA is used to visualize the clusters. The experimental result had demonstrated that DBSCAN is capable of identifying peculiar data points that are deviating from the 'normal' data distribution.

1. Introduction

Weather is highly influential for human life. Anomalies in weather describe conditions that are out of the ordinary and need special attention because they can affect various aspects of human life both socially and economically, including natural disasters. Anomaly detection can aim to get rid of unwanted data (noise, erroneous data, or unwanted data) or to study the anomaly phenomenon itself (unusual but interesting) [1]. The unsupervised anomaly detection method is used when the data object does not have a predefined label, for example as "anomaly" or "normal". This unsupervised method is usually referred to as a clustering problem [2]. Clustering itself is a family of methods in data mining that is useful for finding trends in data groups in a large pile of data. Research has shown that data mining can be used for anomaly detection [3].

In terms of the need for model development, by removing anomaly/outlier data, a better / more general model can be produced [4]. In the case of rain prediction, for example, anomalous removal can have an impact on the formation of a more accurate rain prediction model, making it useful in several fields that are affected by weather such as agriculture, transportation, and so on. The model formation can be made based on historical weather data that has been recorded by meteorological stations spread across various points in Indonesia. This data has been provided by the Meteorology, Climatology, and Geophysical Agency (BMKG) to be accessed by the public for various purposes including research purposes. This study focuses on the detection of anomalies in weather data using a data mining approach. In the absence of an anomaly label, unsupervised learning is used using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method to determine normal and anomalous data

groups. The anomaly detection used in this research is the multivariate type (considering many variables) because it will provide better performance than univariate (based on only one variable).

Research on weather anomaly detection using data mining methods is still very rare. Some of the related studies that have been present are shown in Table 1. In general, anomaly detection can be carried out based on only one variable (univariate outlier detection) or multi-variable (multivariate outlier detection). However, research shows that multivariate outlier detection has better performance than univariate outlier detection for more complex data handling [5]. Multivariate outlier detection has been used in several studies, for example for cases of health claims using the probabilistic programming method [6].

Table 1. Research on weather anomaly detection.

Ref	Anomaly Detection Method	Remark
[7]	*Anomaly Frequency Method* (AFM)	Extreme weather detection
[8]	*Anomaly Frequency Method* (AFM)	Understanding extreme weather using association rule mining
[9]	*Prediction based Outlier Classifier* dan *Pattern-based Outlier Classifier*	Applied to weather data from the IoT sensor

In data mining, there are distance-based outlier detection, clustering-based outlier detection, density-based outlier detection, and depth-based outlier detection [10]. For this study, the density-based clustering method will be used, namely the DBSCAN method. DBSCAN can detect clusters with arbitrary shapes and has few parameters to set. Recent research has used DBSCAN for mapping wildfire-prone areas [11]. Regardless of some research on improving DBSCAN, the original DBSCAN can still perform well [12]. Other studies also show that DBSCAN provides better clustering quality than the PAM and CLARA clustering methods for big data in agriculture [13].

2. Methods

Daily weather data for Tanjung Mas, Semarang City, Indonesia were obtained from the BMKG website. Of the 11 attributes, only 8 attributes are used as shown in Table 2. Some data containing 8888 (unmeasured data) and 9999 (no data / no measurements were taken) were replaced with null. Data cleaning is performed for incomplete data sets. The data is then stored in CSV format and can then be processed using RStudio software. Normalization needs to be done so that none of the attributes are more dominant or less dominant due to differences in measurement scales. The clustering experiment was carried out with the 'dbscan' function in the 'dbscan' package available for R. Eps value will be based on a visual examination of the 'knee' in the K-NN distance plot using the 'kNNdistplot' function in the aforementioned package. KNN is a method commonly used for classification, such as hoaxes classification [14]. The eps and minpts experiment were done several times to get good scores. The final eps and minpts values will be chosen based on the experimental results. The DBSCAN algorithm will produce membership points labeled as cluster members or as noise. Points that are included as noise are anomalous data. While normal data will be covered with a convex hull to show its normal limits. The results of the cluster will be evaluated and understood with the help of Principal Component Analysis (PCA) techniques to understand the characteristics of weather anomalies that are found. PCA is a technique for dimensionality reduction. PCA is also used in face-detection [15] along with Artificial Neural Network.

Table 2. The attributes of the weather data.

Attribute	Data type	Description
Tn	Numeric	Minimum temperature
Tx	Numeric	Maximum temperature
Tavg	Numeric	Average temperature
RH_avg	Numeric	Average Humidity (%)
ss	Numeric	Sun exposure time (hours)
ff_x	Numeric	Maximum wind speed (m/s)
ff_avg	Numeric	Average wind speed (m/s)
RR	Numeric	Rainfall (mm)

2.1. DBSCAN clustering algorithm

Introduced in 1996 [16], DBSCAN uses eps and minpts to determine the cluster. Eps is the maximum distance from a point to evaluate if the other point belongs to the same cluster membership. On the other hand, minpts is the minimum number of points to be considered to determine if a point belongs to the member of the cluster within the radius eps. DBSCAN gives each point a circle with a radius of eps, followed by a membership evaluation for each point included in that circle. Each point will fall into one of three categories, i.e. the core points, border points, or noise points. A point is defined as a core point if it has at least some members (enclosed by the circle) equal to minpts in the radius of eps. A point is defined as a border point when a point is within eps but has some member points that are less than minpts. Meanwhile, if the point is not a core point or border point, then the point is designated as a noise point. The noise point is not a member of any cluster. A cluster is then defined as a membership set that contains a combination of core points surrounded by border points.

3. Results and discussion

The determination of eps for DBSCAN was done by examining the 'knee in the KNN distance plot with k = 5 as shown in Figure 1. Based on this 'knee', eps is set to 0.015. The result of the cluster is shown in Figure 2 and 3 shows the PCA plot with the attribute axes. It shows that Principal Component 1 (PC1) mainly consists of Tx, RH, and Tavg. Whereas PC2 mainly consists of Tn. On the other hand, ff, ff_avg, and ss are the less-prominent components. It is interesting to see that the data formed like stacked eyebrows. This PCA plot shows a trend that for a certain Tn (minimum temperature), the Tx (maximum temperature) and RH_avg (average relative humidity) will lie in a certain range. This will be investigated and explored in future research. Generally, our result showed that anomalous weather is characterized by high humidity and low temperature. However, it is worth noting that an experiment with different data might result in different anomaly characteristics depending on the data used. This research only proposes a method for clustering normal and anomalous data.

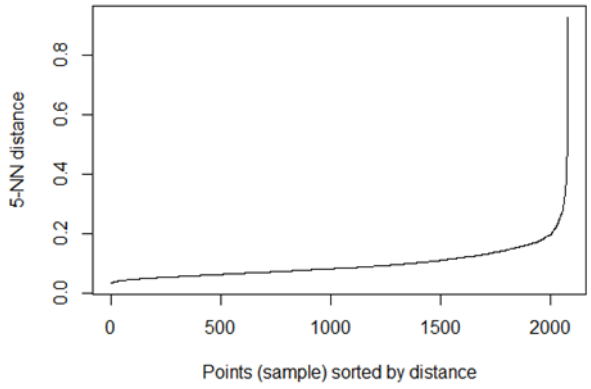

Figure 1. The KNN distance plot.

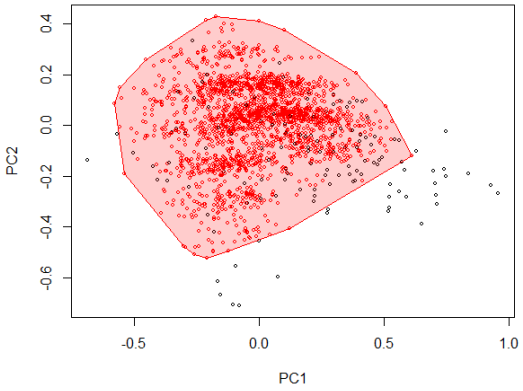

Figure 2. The cluster of 'normal' data (red dots enclosed by red convex hull) and the anomaly/outlier data points (black dots).

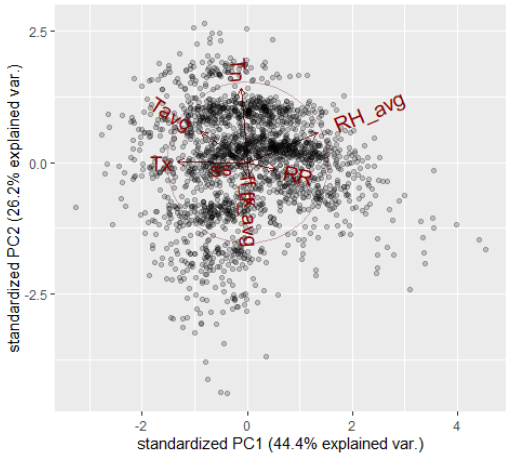

Figure 3. The PCA plot and the attributes axes.

4. Conclusion

We explore the use of the DBSCAN clustering algorithm to identify anomalous weather data from the normal one by considering multiple weather variables. Then, PCA is used to visualize the clusters. The experimental result had demonstrated that DBSCAN is capable of identifying peculiar data points that are deviating from the 'normal' data distribution. In our result, the anomalous weather is characterized by high humidity and low temperature.

Acknowledgments

We thank the Meteorology, Climatology, and Geophysical Agency (BMKG) for providing the weather data.

References

[1] Aggarwal C C 2016 *Outlier analysis second edition* (Switzerland: Springer)
[2] Kaur R and Singh S 2016 A survey of data mining and social network analysis based anomaly detection techniques *Egypt. informatics J.* **17** 199–216
[3] Agrawal S and Agrawal J 2015 Survey on anomaly detection using data mining techniques *Procedia Comput. Sci.* **60** 708–13
[4] Domingues R, Filippone M, Michiardi P and Zouaoui J 2018 A comparative evaluation of outlier detection algorithms: Experiments and analyses *Pattern Recognit.* **74** 406–21
[5] Sunderland K M, Beaton D, Fraser J, Kwan D, McLaughlin P M, Montero-Odasso M, Peltsch A J, Pieruccini-Faria F, Sahlas D J, Swartz R H and others 2019 The utility of multivariate outlier detection techniques for data quality evaluation in large studies: an application within the ONDRI project *BMC Med. Res. Methodol.* **19** 102
[6] Bauder R A and Khoshgoftaar T M 2017 Multivariate outlier detection in medicare claims payments applying probabilistic programming methods *Heal. Serv. Outcomes Res. Methodol.* **17** 256–89
[7] Piruthevi C and Selvi C S K 2017 Filtering of anomalous weather events and tracing their behavior *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* pp 1–5
[8] Piruthevi C and Selvi C S K 2017 Filtering of anomalous weather events over the region of Tamil Nadu *2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)* pp 1–7

[9] Saneja B and Rani R 2018 A Hybrid Approach for Outlier Detection in Weather Sensor Data *2018 IEEE 8th International Advance Computing Conference (IACC)* pp 321–6
[10] Bansal R, Gaur N and Singh S N 2016 Outlier detection: applications and techniques in data mining *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)* pp 373–7
[11] Anwar M T, Hadikurniawati W, Winarno E and Supriyanto A 2019 Wildfire Risk Map Based on DBSCAN Clustering and Cluster Density Evaluation *Adv. Sustain. Sci. Eng. Technol.* **1**(1)
[12] Schubert E, Sander J, Ester M, Kriegel H P and Xu X 2017 DBSCAN revisited, revisited: why and how you should (still) use DBSCAN *ACM Trans. Database Syst.* **42** 1–21
[13] Majumdar J, Naraseeyappa S and Ankalaki S 2017 Analysis of agriculture data using data mining techniques: application of big data *J. Big data* **4** 20
[14] Zuliarso E, Anwar M T, Hadiono K and Chasanah I 2020 Detecting Hoaxes in Indonesian News Using TF/TDM and K Nearest Neighbor *IOP Conference Series: Materials Science and Engineering* **835** 12036
[15] Winarno E, Al Amin I H, Februariyanti H, Adi P W, Hadikurniawati W and Anwar M T 2019 Attendance System Based on Face Recognition System Using CNN-PCA Method and Real-time Camera *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* pp 301–4
[16] Ester M, Kriegel H-P, Sander J, Xu X and others 1996 A density-based algorithm for discovering clusters in large spatial databases with noise *Kdd* **96** 226–31