International Journal of Information System Modeling and Design
Volume 13 • Issue 3

AI for Health-Related Data Modeling:
DCN Application Analysis

Na Cheng, Jilin Institute of Physical Education, China*

ABSTRACT

Data modeling of health-related data from Data Center (DC) has positive effects for health monitoring, disease prevention, and healthcare research. However, health-related data has the characteristics of huge, high-dimensional, and non-normalized, which are not beneficial to direct analysis, so data needs to be preprocessed before data modeling. This paper focuses on the features of health-related data, and outlier detection during data preprocessing is studied. Meanwhile, the authors propose an improved algorithm for health-related data-based outlier detection. The experimental results reveal that the proposed outlier detection algorithm has a smaller running time, and more outliers are detected compared to three baselines. In addition, local importance-based random forest feature selection algorithm is proposed to measure the importance of each feature. The experimental results indicate that the proposed algorithm can select optimal feature subset to apply health-related data.

KEYWORDS

Application Analysis, Data Modeling, DCN, Feature Selection, Health, K-Means, Local Importance, Outlier Detection

1. INTRODUCTION

Since last few years, due to the emerging technologies such as cloud computing, big data, and Internet of Things (IoT) (Joshi et al., 2021; Muniswamaiah et al., 2019), the volume of data is increasing day by day. The rapid growth of web applications including search engine, online shopping, and cloud computing is putting forward severe requirements on the underlying infrastructure in terms of computing, storage, and networking. In order to meet the storage and processing needs of large amounts of data, Data Center (DC) has become an indispensable information platform, which is responsible for the management and maintenance of massive computing and storage systems. Internet companies like Microsoft, Google, Amazon, Facebook, and Alibaba have built high-performance data centers around the world. These data centers connect servers and network switches over network to meet the needs of high-speed computing and massive storage in a more convenient way. While Data Center Network (DCN) plays a crucial role in data center by connecting all the data center resources together (Chen et al., 2021).

Machine Learning (ML) is a very successful approach of Artificial Intelligence (AI) (Di Mitri et al., 2017; Phellan et al., 2021), which is the core of AI, and it is also a form of AI in which the computer learns how to complete a task by itself. ML can help machines to take right decisions and smart actions in real time without human intervention. There are two common ML models that

DOI: 10.4018/IJISMD.300780 *Corresponding Author

are supervised learning model and unsupervised learning model. ML has been around for a while which has grown at a high speed in recent years. In future, ML will be one of the best solutions for analyzing large amounts of data. If handled right, ML could change the way humans live more than any technology that ever existed.

As more and more people begin to connect to the Internet, data from DC is increasing, but health-related data is what we are concerned about. Health has always been part of our whole way of life. Every part of our life relies on having good health. Living a healthy lifestyle can help prevent chronic diseases and long-term illnesses. The importance of good health in our life is undoubtedly great. Accordingly, the main contributions of this paper are summarized as follows. (i) Combining voting strategy based global outlier detection with K-means based nearest-furthest neighbors search, an improved algorithm for health-related data based outlier detection algorithm is proposed. (ii) We propose local importance based random forest feature selection algorithm to measure the importance of each feature.

The remaining of this paper is organized as follows. Section 2 reviews the related work. In Section 3, two algorithms are proposed in terms of data preprocessing. The experimental results are shown in Section 4. Section 5 concludes this paper.

2. RELATED WORK

Many strategies on outlier detection have been proposed. In (Wang et al., 2021), R-tree based outlier detection algorithm was proposed, which can effectively support single query and multiple query processing. In (Shao et al., 2021), an advanced fast density peak outlier detection algorithm based on the characteristics of big data was proposed to avoid the clustering process and reduce the running time of the cluster-based outlier detection algorithm. In (Zhou et al., 2019), a novel outlier detection as was proposed to integrate the local density with the global distance seamlessly. In the proposed method, an integrated outlier factor was used to measure the detecting accuracy. In (Erkus & Purutcuoglu, 2021), a new non-parametric outlier detection technique was suggested which was based on the frequency-domain. In (Du et al., 2021), a simple scheme outlier detection tree based on entropy was described. Each data object was identified as an outlier or a normal one using the if-then rules in the tree. Furthermore, an advanced outlier detection algorithm was proposed to achieve both high detection accuracy and low time complexity. In (Wahid & Annavarapu, 2021), an unsupervised density-based outlier detection algorithm was presented to deal with parameter selection problem. In (Lai et al., 2020), a privacy-preserving outlier detection protocol for incremental data sets was presented. The protocol decomposed the outlier detection algorithm into several phases and recognized the necessary cryptographic operations in each phase. It realized several cryptographic modules via efficient and interchangeable protocols to support the above cryptographic operations and composed them in the overall protocol to enable outlier detection over encrypted data sets. In (Goh et al., 2020), a novel combinatory-distance-based method capable of high accuracy outlier detection named as the sorted distance divergence point was introduced. Moreover, the introduced distance function and outlier percentage allowed clear labelling of inliers and outliers cloud points. In (Wang & Mao, 2019), an outlier detection scheme was proposed that can be directly used for either process monitoring or process control. Based on traditional Gaussian process regression, authors developed several detection algorithms, of which the mean function, covariance function, likelihood function and inference method were specially devised. In (Cai et al., 2019), a two-phase pattern-based outlier detection approach was proposed for effectively detecting the implicit outliers from a weighted data stream, in which the maximal frequent patterns were used instead of the frequent patterns to accelerate the process of outlier detection.

Some approaches on feature selection have also been proposed. In (Munirathinam & Ranganadhan, 2020), a novel feature selection algorithm based on Chebyshev distance-outlier detection model was proposed. In (Dornaika, 2021), three schemes for the joint feature and instance selection were

proposed. The first was a wrapper technique while the two remaining ones were filter approaches. In (Bakhshandeh et al., 2020), a filer-based feature selection algorithm using graph technique was proposed for reducing the dimension of dataset. In (Morillo-Salas et al., 2021), a methodology to distribute feature selection processes based on selecting relevant and discarding irrelevant features was proposed. In (Ben Brahim, 2021), a filter method was proposed that improved stability of feature selection while preserving an optimal predictive accuracy and without increasing the complexity of the feature selection algorithms. In (Sadeg et al., 2020), the use of feature selection to speed-up the search process of bee swarm optimisation meta-heuristic in solving the MaxSAT problem was investigated. In (Alarifi et al., 2020), a novel big data and machine learning technique were introduced for evaluating sentiment analysis processes. From the cleaned sentiment data, effective features were selected using a greedy approach that selected optimal features processed by an optimal classifier called cat swarm optimization-based long short-term memory neural network. The classifiers analyzed sentiment-related features according to cat behavior, minimizing error rate while examining features. In (Jijesh et al., 2021), a supervised learning based decision support system for multi sensor healthcare data from wireless body sensor networks was proposed. Data fusion ensemble scheme was developed along with medical data which was obtained from body sensor networks. Ensemble classifier was taken the fusion data as an input for heart disease prediction. Feature selection was done by the squirrel search algorithm which was used to remove the irrelevant features. In (Qiu, 2020), a novel hybrid two-stage feature selection method based on differential evolution was proposed. In the first stage, a cluster validity index named DB index was employed to evaluate the feature subset and the wrapper approach was used in the second stage to improve the classification accuracy of the feature subset. In (Jo et al., 2019), a method to improve the performance of minimum redundancy maximum relevance feature selection was proposed.

3. DATA PREPROCESSING

3.1 Outlier Detection

Data preprocessing plays a pivotal role in data modeling, which enhances the quality of raw data for further processing. Health-related data from DC is often incomplete, full with noise, and inconsistent. It is difficult to perform data modeling by using useless data, so data preprocessing is necessary. There are a series of steps during data preprocessing such as data cleaning, data integration, data reduction, and data transformation.

Typically, health-related data is often taken from multiple sources, which is duplicated, inaccurate, redundant, incomplete, or incorrect. Due to the low quality of health-related data from DC, it is unfavourable for data modeling.

3.1.1 Voting Strategy Based Global Outlier Detection

There is a large amount of inconsistent data in health-related dataset, which is called outliers. The existence of outliers for many reasons such as different measurements, or may be errors in data entry, so it is difficult to detect outliers by using traditional density based outlier detection algorithm. Outliers are usually cleansed as anomalous data during data preprocessing. The mining of anomalous data is outlier detection. Unrelated information in health-related data can be removed through outlier detection in order to improve the performance of data analysis.

The most common methods for outlier detection are statistical-based, distance-based, density-based, deviation-based, and clustering-based. The improved outlier detection algorithm based on health-related data is divided into two parts in this paper: voting strategy based global outlier detection and K-means based nearest-furthest neighbors search.

The thought of voting strategy based global outlier detection improved algorithm proposed in this paper is: the difference between the anomalous data deviating from the dataset and the normal

data is greater than the difference between the normal data. A voting strategy based global outlier detection improved algorithm allows each sample point from health-related dataset to vote for the samples with the greatest difference. After voting by all sample points, the points with more votes are significantly different from samples, and they are more likely to become outliers.

3.1.2 K-Means Based Nearest-Furthest Neighbors Search

In density based outlier detection algorithm, nearest neighbors and furthest neighbors of each sample point are needed to be calculated. In addition, this calculation process requires the euclidean distance between the sample point and the others. It will also consume time due to the increasing of dataset. The introduction of K-means algorithm can greatly reduce the running time of the algorithm (Gu et al., 2019). In clustering process, there are some additional information for each sample point, which are the cluster number of the sample point and the distance between the sample point and the centroid of cluster. Similarly, there is also additional information for the centroid of cluster. Calculating the furthest distance between a point in the cluster and the centroid of cluster, which is taken as the radius r of the cluster. The strategy is summarized as follows:

1. The process of searching for the nearest neighbors of any point P:
   Step1: Adding the points in the cluster of point P to the search range.
   Step2: Calculating the distance between given point and in turn. Assuming that given point P is in cluster A, d(M)+r(M) is the distance between point P and the centroid of cluster M plus the radius of cluster M, which means that the furthest possible distance between point P and point in cluster M. While d(N)-r(N) is the difference between the distance from point P to the centroid of cluster N and the radius of cluster N, which means that the nearest possible distance between point P and point in cluster M. If $d(M)+r(M) > d(N)-r(N)$, when searching for the k nearest neighbors of point P, the points in cluster N is added to the search range.
   Step3: Calculating by step2 with other clusters one after the other to determine the final search range.
   Step4: Searching for k points nearest to the given point in the search range, and adding them to the related k nearest neighbors.
2. The process of searching for the furthest neighbors of point P:
   Step1: Calculating the distance between the given point and centroid of other clusters in turn.
   Step2: If $d(N) + r(N) > d(M) - r(M)$, then points in cluster B are added to the search range for the n furthest neighbors. The algorithm continues to traverse other clusters, and determine the search range of the n furthest neighbors of the given point, then search the furthest n points from the given point in the search range, finally increase the votes of the n points by one.
   Step3: Calculating by step2 with other clusters one after the other to determine the final search range.
   Step4: Searching for n points furthest to the given point in the search range, and adding them to the related n furthest neighbors.

The number of cluster should be set to a larger value so as to ensure that the volume of data in each cluster is kept within an appropriate range, which can effectively reduce the search range and improve the performance of the algorithm. The improved algorithm greatly reduces the search range when searching for the nearest neighbors and furthest neighbors.

3.1.3 An Improved Algorithm for Health-Related Data Based Outlier Detection

Based on the voting strategy based global outlier detection and K-means based nearest-furthest neighbors search, the process of the improved algorithm is summarized as follows:

Step1: K-means clustering is performed on the health-related data to cluster the data into clusters of different sizes.

Step2: Calculating the euclidean distance between any point X_i and other point Y_j in the same cluster. Finding k points nearest to X_i, and calculating the local outlier factor (LOF) of this point. LoF is the average of the ratio of the local reachability density of a point in the neighbors of point P to that of point P.

Step3: Marking m points that furthest from any point X_i: X_i is in cluster A, and searching cluster B furthest from cluster A. Calculating the distance between X_i and any point in cluster B, then searching n points furthest from point X_i. Finally, adding n points to the furthest neighbors of X_i, and adding one to the number of votes for each point in the furthest neighbors.

Step4: Local outliers and global outliers are identified according to the threshold of LOF and global outliers.

3.2 Local Importance Based Random Forest Feature Selection Algorithm

Features are often called as attributes, properties, or dimensions. Data objects are described by some features. In ML, feature selection, also known as variable selection, is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. Feature selection is removing the irrelevant data from health-related dataset, then using the feature selection algorithm to measure features.

It is unnecessary to process some redundant or irrelevant features in complicated health-related data, which also reduces the accuracy of data analysis. In health-related data, some attributes like height, weight, and body mass index (BMI) may cause the redundancy because BMI can be expressed by height and weight. In addition, a large number of features in health-related data with redundant information will easily result in long time for feature analysis and model training. Therefore, it is necessary to perform feature selection during data preprocessing.

Random forest (RF) is a supervised learning algorithm that randomly constructs multiple independent decision trees to form one forest. The traditional random forest based feature selection algorithm has the following problems.

1. In the process of the construction of the random forest, the feature subset will be selected randomly for the construction of each decision tree, but there is no guarantee that all features will participate in the construction of the random forest after the completion of the construction of all decision trees. When a feature is not selected for the construction of a decision tree, it cannot be selected into the optimal feature subset even if it is high in importance.
2. When calculating the importance of each feature, the traditional algorithm uses the mean value of the error change of each tree without considering the local feature priority order. However, the importance of each attribute is not the same in each tree, which has been reflected in the process of tree construction. It is possible that the feature with weaker classification ability will bring greater classification error, that is, the feature is more important only in the related decision tree from global perspective, but the order of feature importance in each decision tree is ignored. This will bring deviation to the final calculation of feature importance and affect the selection of feature subset. According to the above, we propose local importance based random forest feature selection algorithm, the improvements are summarized as follows:
    a. The importance cannot be calculated because some features are not selected in the tree construction process. The tree construction can be stopped in case of each feature is selected at least twice. Otherwise, the features can be randomly selected to construct a new decision tree.
    b. When the importance is calculated, the impact factor of local feature order is introduced to balance the error caused by importance calculation. In the process of constructing the random forest, the decision tree used in this paper is C4.5, and the method of selecting split nodes is information gain ratio.

The feature with larger information gain ratio also has stronger classification ability, and the feature importance is higher in local feature subset. We can take the information gain ratio of each feature in the decision tree as the impact factor of local priority. The feature with higher value plays a better role in decision tree, and it also has a stronger local importance.

This section proposes local importance based random forest feature selection algorithm. When the importance is calculated, we introduce local importance of each feature to reflect the ranking of features in feature selection.

4. EXPERIMENTAL AND RESULTS ANALYSIS

4.1 Data Source

We select health-related dataset with the size of 5000, 10000, 20000, and 50000 randomly, and experiments are performed in selected dataset in order to demonstrate that the proposed outlier detection algorithm is more efficient in running time. We choose relative kernel density-based outlier score (RKDOS) (Wahid & Rao, 2020), influenced outlierness (INFLO) (Zhou et al., 2019), and local distance-based outlier detection factor (LDOF) (Radovanović et al., 2015) for comparison.

There are 20 features in health-related data, and 14 features are left after selecting related rules. We perform feature selection for 14 features, and the abbreviations of features used in this paper are listed in Table 1. We choose clustering-based feature selection (CBFS) (Mao et al., 2020), adaptive multi-population genetic algorithm (AMGA) (Shukla, 2020), and local importance based random forest feature selection algorithm proposed in this paper in feature importance for comparison.

4.2 Comparison Analysis

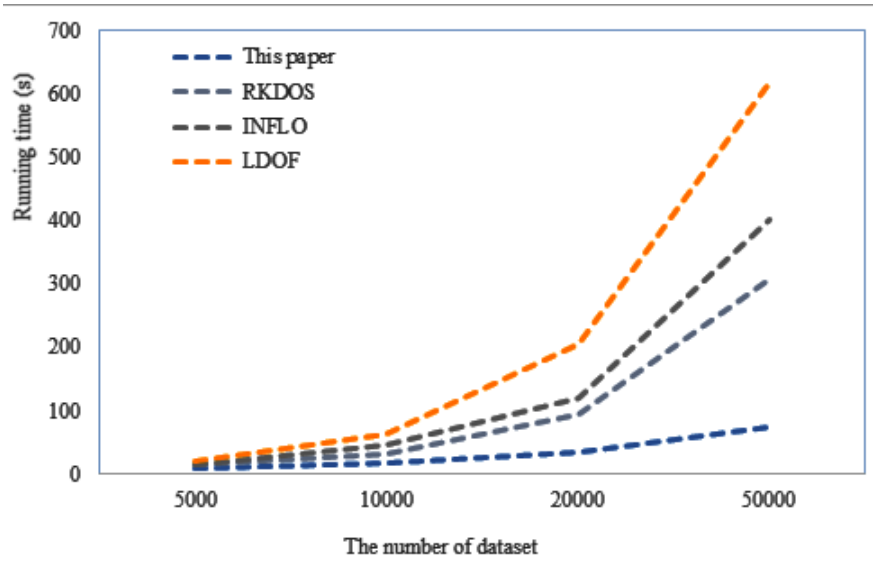
Figure 1 shows that with the increasing of dataset, the running time curves of the comparison algorithms are going to get steeper. While the running time curve of the proposed algorithm in this

Table 1. Description of features

Description	Abbreviation
BMI	BMI
blood pressure	BP
takeout	TO
exercise	EXC
overtime	OT
drink water	DW
stay up late	SUL
alcoholism	ACHL
smoking	SMK
obesity	OBS
sleeping	SLP
snacks	SNK
aggressive driving	AD
genetic disorder	GD

Note. BMI=weight/height^2

Figure 1. The comparison of running time



paper is much smoother, which means that the proposed has a good performance in running time. As can be seen from Figure 2, there are more outliers detected with the increasing of dataset.

In this paper, 14 features are performed for feature selection. After several iterations, there are 7 remaining feature items, and the importance of each feature is shown in Table 2.

Moreover, it can be seen from Table 2 that the feature importance of the algorithm proposed in this paper are generally decrease, and the ranking of each feature importance in three algorithms is not the same. Figure 3 shows the importance comparison in the first iteration. If the top 11 feature importance are selected as the new feature sets, then the feature items which need to be eliminated

Figure 2. The number of outliers detected

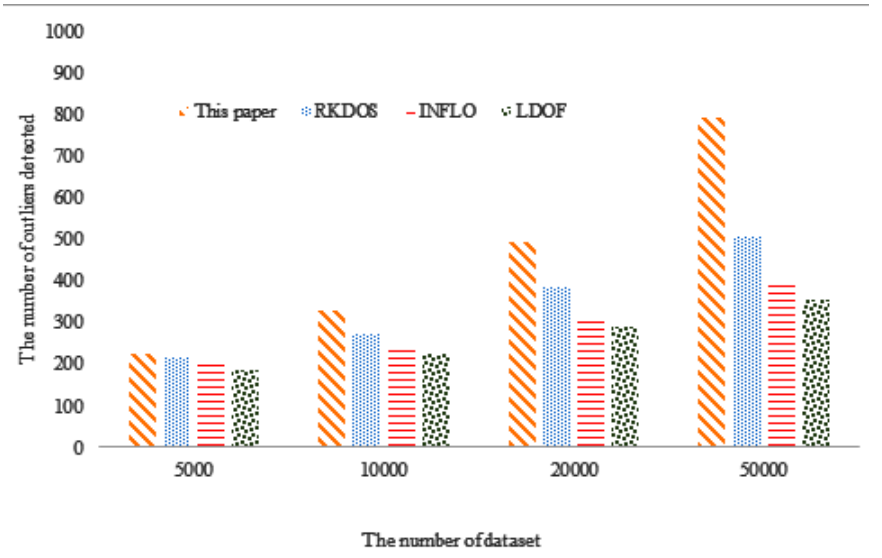
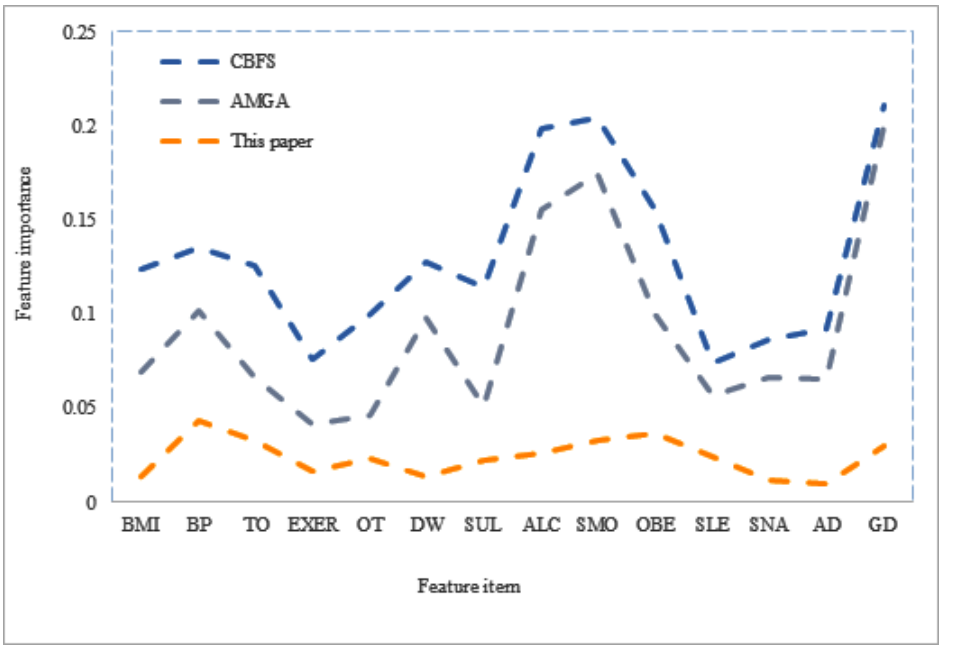


Table 2. The importance of each feature

Feature	1st iteration			5th iteration			10th iteration		
	CBFS	AMGA	This paper	CBFS	AMGA	This paper	CBFS	AMGA	This paper
BMI	0.1242	0.069	0.0136	-	-	-	-	-	-
BP	0.1357	0.1021	0.0431	0.1816	0.0894	0.0453	0.1546	0.0712	0.0189
TO	0.1255	0.0667	0.0325	0.1421	0.0749	0.0626	0.1328	0.0844	0.0176
EXC	0.0755	0.0412	0.0163	0.1126	-	0.0301	-	-	-
OT	0.1002	0.0458	0.0237	-	-	-	-	-	-
DW	0.1278	0.0983	0.0135	-	-	-	-	-	-
SUL	0.1146	0.0515	0.0224	0.2314	0.1842	0.0817	0.1431	0.032	0.0161
ACHL	0.1986	0.1553	0.0266	0.2105	0.0987	0.0663	0.1327	0.0456	0.0202
SMK	0.2041	0.1745	0.0328	0.1367	0.12	0.0418	0.1653	0.074	0.0316
OBS	0.1553	0.0996	0.0366	0.1864	0.1422	0.0726	-	-	0.0199
SLP	0.0744	0.057	0.0241	-	-	0.0388	-	-	-
SNK	0.0869	0.0662	0.012	-	-	-	-	-	-
AD	0.0924	0.0656	0.0097	0.0818	0.0201	0.0244	-	-	-
GD	0.2113	0.2008	0.0304	0.1206	0.1143	0.0584	0.2416	0.1986	0.0506

Figure 3. The comparison of feature importance



are EXC, OT, and SLP in compared algorithms. In the proposed algorithm, the feature items which need to be eliminated are AD, SNK, and DW. After 10 iterations of the algorithm, the remained feature items are GD, SMK, ACHL, OBS, BP, TO, and SUL. In summary, the algorithm proposed in this paper can optimize the problem of ignoring the local importance in the process of random forest feature selection, and make the final selection of feature subset more optimal.

5. CONCLUSION

Health is above everything else. Without health nothing can be done, so data modeling of health-related data from DC is necessary. Data preprocessing plays a pivotal role in data modeling. This paper studies outlier detection in data preprocessing. There are outliers in health-related data, for the low efficiency of density based outlier detection algorithm and the drawback of insensitive to global outlier detection, combining K-means clustering and voting strategy, an improved algorithm for health-related data based outlier detection algorithm is proposed. After experimental verification, the algorithm proposed in this paper has good performance in outlier detection as well as running time. For the feature selection problem in health-related data, local importance based random forest feature selection algorithm is proposed. It demonstrates that the proposed algorithm can optimize the problem of ignoring the local importance in the process of random forest feature selection through the experiment. In addition, when the initial feature subset is large in proposed algorithm, a large number of decision trees may be generated in order to meet the terminate condition, which reduces the performance of the algorithm. Therefore, it is necessary to optimize the performance of the algorithm.

ACKNOWLEDGMENT

This paper is supported by Jilin Province 13th Five-Year Educational Science Planning Project in 2020 (Granted Number GH20374).

FUNDING AGENCY

Publisher has waived the Open Access publishing fee.

REFERENCES

Alarifi, A., Tolba, A., Al-Makhadmeh, Z., & Said, W. (2020). A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks. *The Journal of Supercomputing*, *76*(6), 4414–4429. doi:10.1007/s11227-018-2398-2

Bakhshandeh, S., Azmi, R., & Teshnehlab, M. (2020). Symmetric uncertainty class-feature association map for feature selection in microarray dataset. *International Journal of Machine Learning and Cybernetics*, *11*(1), 15–32. doi:10.1007/s13042-019-00932-7

Ben Brahim, A. (2021). Stable feature selection based on instance learning, redundancy elimination and efficient subset fusion. *Neural Computing & Applications*, *33*(4), 1221–1232. doi:10.1007/s00521-020-04971-y

Cai, S., Li, Q., Li, S., Yuan, G., & Sun, R. (2019). WMFP-Outlier: An Efficient Maximal Frequent-Pattern-Based Outlier Detection Approach for Weighted Data Streams. *Information Technology and Control*, *48*(4), 505–521. doi:10.5755/j01.itc.48.4.22176

Chen, G., Cheng, B., & Wang, D. (2021). Constructing Completely Independent Spanning Trees in Data Center Network Based on Augmented Cube. *IEEE Transactions on Parallel and Distributed Systems*, *32*(3), 665–673. doi:10.1109/TPDS.2020.3029654

Di Mitri, D., Scheffel, M., & Drachsler, H. (2017). Learning Pulse: A machine learning approach for predicting performance in self-regulated learning using multimodal data. Seventh International Learning Analytics & Knowledge Conference. doi:10.1145/3027385.3027447

Dornaika, F. (2021). Joint feature and instance selection using manifold data criteria: Application to image classification. *Artificial Intelligence Review*, *54*(3), 1735–1765. doi:10.1007/s10462-020-09889-4

Du, H., Ye, Q., Sun, Z., Liu, C., & Xu, W. (2021). FAST-ODT: A Lightweight Outlier Detection Scheme for Categorical Data Sets. *IEEE Transactions on Network Science and Engineering*, *8*(1), 13–24. doi:10.1109/TNSE.2020.3022869

Erkus, E. C., & Purutcuoglu, V. (2021). Outlier detection and quasi-periodicity optimization algorithm: Frequency domain based outlier detection (FOD). *European Journal of Operational Research*, *291*(2), 560–574. doi:10.1016/j.ejor.2020.01.014

Goh, M. J. S., Chiew, Y. S., & Foo, J. J. (2020). Outlier percentage estimation for shape- and parameter-independent outlier detection. *IET Image Processing*, *14*(14), 3414–3421. doi:10.1049/iet-ipr.2020.0334

Gu, Y., Li, K., Guo, Z., & Wang, Y. (2019). Semi-supervised k-means ddos detection method using hybrid feature selection algorithm. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 64351–64365. doi:10.1109/ACCESS.2019.2917532

Jijesh, J. J., Shivashankar, , & Keshavamurthy, . (2021). Shivashankar, Keshavamurthy, A Supervised Learning Based Decision Support System for Multi-Sensor Healthcare Data from Wireless Body Sensor Networks. *Wireless Personal Communications*, *116*(3), 1795–1813. doi:10.1007/s11277-020-07762-9

Jo, I., Lee, S., & Oh, S. (2019). Improved Measures of Redundancy and Relevance for mRMR Feature Selection. *Computers*, *8*(2), 42–55. doi:10.3390/computers8020042

Joshi, M., Budhani, S., & Tewari, N. (2021). Analytical Review of Data Security in Cloud Computing. *Proceedings of 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, 362-366. doi:10.1109/ICIEM51511.2021.9445355

Lai, S., Yuan, X., Sakzad, A., Salehi, M., Liu, J. K., & Liu, D. (2020). Enabling efficient privacy-assured outlier detection over encrypted incremental data sets. *IEEE Internet of Things Journal*, *7*(4), 2651–2662. doi:10.1109/JIOT.2019.2949374

Mao, J., Hu, Y., Jiang, D., Wei, T., & Shen, F. (2020). CBFS: A Clustering-Based Feature Selection Mechanism for Network Anomaly Detection. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 116216–116225. doi:10.1109/ACCESS.2020.3004699

Morillo-Salas, J. L., Bolon-Canedo, V., & Alonso-Betanzos, A. (2021). Dealing with heterogeneity in the context of distributed feature selection for classification. *Knowledge and Information Systems*, *63*(1), 233–276. doi:10.1007/s10115-020-01526-4

Munirathinam, D. R., & Ranganadhan, M. (2020). A new improved filter-based feature selection model for high-dimensional data. *The Journal of Supercomputing*, *76*(8), 5745–5762. doi:10.1007/s11227-019-02975-7

Muniswamaiah, M., Agerwala, T., & Tappert, C. (2019). Big data in cloud computing review and opportunities. *International Journal of Computer Science and Information Technologies*, *11*(4), 43–57. doi:10.5121/ijcsit.2019.11404

Phellan, R., Hachem, B., Clin, J., Mac-Thiong, J.-M., & Duong, L. (2021). Real-time biomechanics using the finite element method and machine learning: Review and perspective. *Medical Physics*, *48*(1), 7–18. doi:10.1002/mp.14602 PMID:33222226

Qiu, C. (2020). A hybrid two-stage feature selection method based on differential evolution. *Journal of Intelligent & Fuzzy Systems*, *39*(1), 871–884. doi:10.3233/JIFS-191765

Radovanović, M., Nanopoulos, A., & Ivanović, M. (2015). Reverse nearest neighbors in unsupervised distance-based outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, *27*(5), 1369–1382. doi:10.1109/TKDE.2014.2365790

Sadeg, S., Hamdad, L., Chettab, H., Benatchba, K., Habbas, Z., & Kechadi, M.-T. (2020). Feature selection based bee swarm meta-heuristic approach for combinatorial optimisation problems: A case-study on MaxSAT. *Memetic Computing*, *12*(4), 283–298. doi:10.1007/s12293-020-00310-9

Shao, M., Qi, D., & Xue, H. (2021). Big data outlier detection model based on improved density peak algorithm. *Journal of Intelligent & Fuzzy Systems*, *40*(4), 6185–6194. doi:10.3233/JIFS-189456

Shukla, A. K. (2020). Multi-population adaptive genetic algorithm for selection of microarray biomarkers. *Neural Computing & Applications*, *32*(15), 11897–11918. doi:10.1007/s00521-019-04671-2

Wahid, A., & Annavarapu, C. S. R. (2021). NaNOD: A natural neighbour-based outlier detection algorithm. *Neural Computing & Applications*, *33*(6), 2107–2123. doi:10.1007/s00521-020-05068-2

Wahid, A., & Rao, A. C. S. (2020). RKDOS: A Relative Kernel Density-based Outlier Score. *IETE Technical Review*, *37*(5), 441–452. doi:10.1080/02564602.2019.1647804

Wang, B., & Mao, Z. (2019). Outlier detection based on Gaussian process with application to industrial processes. *Applied Soft Computing*, *76*, 505–516. doi:10.1016/j.asoc.2018.12.029

Wang, X., Li, J., Bai, M., & Ma, Q. (2021). RODA: A Fast Outlier Detection Algorithm Supporting Multi-Queries. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 43271–43284. doi:10.1109/ACCESS.2021.3058660

Zhou, H., Liu, H., Zhang, Y., & Zhang, Y. (2019). An outlier detection algorithm based on an integrated outlier factor. *Intelligent Data Analysis*, *23*(5), 975–990. doi:10.3233/IDA-184227