International Scientific Journal of Engineering and Management ISSN: 2583-6129
Volume: 03 Issue: 04 | April – 2024 DOI: 10.55041/ISJEM01499
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

OUTLIER DETECTION

C.Vasuki[1], S.Boobathiraja2, M.Keerthana[3], S.Lohit[4], C.Sowmiya[5]

[1] Assistant Professor in Information Technology, Nandha Engineering College, Perundurai-638052, Erode District, Tamilnadu, India.

[2,3,4,5] Students of Information Technology, Nandha Engineering College, Perundurai-638052, Erode District, Tamilnadu, India.

ABSTRACT

To find observations that differ considerably from the bulk of the data points, outlier detection is an essential task in data analysis. To put it more simply, outliers are individual data points that stand out from the rest of the dataset. the Iris dataset, a machine learning benchmark, is used for outlier detection. The Rank SVM method, which is typically utilized for ranking jobs but has been modified for outlier identification, is used to find outliers. Standardizing the features pre-processes the dataset, which consists of measurements of iris flower sepal and petal diameters. The standardised dataset is used to train a Rank SVM model. According to the model's categorization, outliers are anticipated to be categorized as -1 and inliers as 1. The study shows the indices of outliers found in the Iris dataset and sheds light on how well Rank SVM works for outlier identification tasks.

Keywords:outlier detection, iris dataset, human behavior, machine learning

1. INTRODUCTION

Finding outlier patterns is an essential part of data research, particularly in industries like cybersecurity, healthcare, and finance. Anomalous data points, often known as outliers, can provide important information about mistakes, unforeseen occurrences, and even fraud. Conventional approaches to outlier detection frequently depend on machine learning algorithms that have been trained on historical data or on predetermined statistical criteria. Nevertheless, these methods might not be able to identify new abnormalities or adjust to changing data distributions. In order to tackle this issue, scholars have suggested utilizing artificially produced searching conditions based on queries. This novel method allows for more accurate and adaptable outlier detection by dynamically creating search criteria based on the properties of the data being evaluated. The purpose of this study is to investigate the idea of query-based artificially produced searching conditions and their possible uses in outlier identification. We will address implementation methodologies and emphasize the advantages of this approach.

1.1 HUMAN BEHAVIOUR

Human conduct is the wide range of attitudes, behaviours, and feelings that people show in different settings and circumstances, both individually and in groups. A complex interaction of biological, psychological, social, cultural, and environmental variables influences it. In disciplines like psychology, sociology, anthropology, economics, and neuroscience, where decision-making processes, social interactions, communication patterns, and societal dynamics are studied, it is essential to observe and comprehend human behavior. Human behavior may range from innate responses to intentional behaviours, and it can also be both predictable and unpredictable. It frequently includes intricate cognitive functions that are impacted by both internal and external inputs, such as perception, attention, memory, reasoning, and problem-solving. Human behavior is also shaped and modulated by individual variations, society standards, and cultural norms, producing a variety of manifestations and results.

1.2 OUTLIER DETECTION

Finding data points that substantially differ from the bulk of the dataset is the goal of outlier identification, an essential part of data analysis. These abnormalities, also referred to as outliers, can occur for a number of causes, including measurement mistakes, noise from experiments, or truly uncommon occurrences. Finding the divergence of each data point from the dataset's expected or typical behavior is the first step in the outlier identification procedure. Conventional ways for detecting outliers use statistical methods like box plots, z-score analysis, and clustering algorithms like k-means. More sophisticated techniques use machine learning methods to identify anomalies in high-dimensional datasets or ones with intricate patterns, such as neural networks, isolated forests, and one-class SVM. Applications for outlier detection can be found in a wide range of fields, such as industrial operations, healthcare, and finance. In finance, it is used to detect fraudulent transactions and rare diseases. In healthcare, it is used to diagnose abnormal medical conditions and equipment failures. The integrity and dependability of analytical models and systems are preserved, decision-making is strengthened, and data quality is improved through effective outlier identification.

2. LITERATURE SURVEY

In this research, Xinfeng Zhang [1] et al. claim that the mobility of crowds in surveillance footage is similar to the thermal motion of elementary particles. Motivated by this, we provide Boltzmann Entropy as a tool for quantifying crowd motion in an optical flow field and identifying anomalous collective behaviours. Consequently, it is possible to depict the collective crowd movement pattern as a time series. We discovered that an anomalous behavior by the majority of people results in a sharp rise in the entropy number. Therefore, without the need for machine learning, a threshold may be applied to the time series to quickly and easily identify instances of unusual crowd disturbance. Comparing the experimental results to state-of-the-art techniques, promising performance is shown. The technology operates with extreme accuracy in real time. When emerging catastrophes like natural disasters or terrorist attacks occur, crowd disturbances might serve as warning signs since they should lead individuals to behave strangely. Here, modeling crowd motion to account for various motion patterns is the main challenge for crowd motion monitoring. Conventional approaches handle each individual as a separate item for further investigation. However, because of the high person-to-object density in crowded environments, object recognition and tracking is not feasible. Modeling people's collective mobility at the particle level—that is, using pixels, picture patches, or local 3D cuboids—has become more popular recently.

This research considers hourly PM10 readings from 22 monitoring sites located in the French regions of Haute-Normandie and Basse-Normandie, as well as in nearby areas, as suggested by MICHEL BOBBIA [2] et al. Every monitoring station that is taken into consideration is either a rural or urban backdrop station. The main objective of the article is to identify statistically any hourly PM10 concentration outliers from a geographical perspective. The overall plan compares the actual measurement with a reliable spatial forecast using a jackknife-style methodology. Two geographical forecasts are taken into consideration: the first is based on the weighted concentrations of the nearest nearby stations, which is based on the median of their concentrations; the second, instead of using more conventional pseudo-innovations, is based on kriging increments. The two techniques are completely integrated by Air in the Measurements Quality Control process and applied to the PM10 monitoring network in Normandy. To demonstrate and contrast the two approaches, some numerical findings based on recent data from January 1, 2013, to May 31, 2013, are given. An governmental group in France keeps an eye on the state of the air in every region. Air quality in Normandy (which is divided into two areas) is monitored by Air Normand, which is situated in Rouen, and Air C.O.M., or Air COM, which is based in Caen. Apart from these principal duties, they also have the responsibility of educating the public about air quality.

Yu [3] et al. Because science problems are so complicated, researchers are switching from doing solitary studies to doing collaborative studies. Academic teams with leaders are built to accelerate information exchange and issue solving in order to obtain greater performance and reputations. Investigating team-based problems is important given the growing interest in information exploration in large-scale academic data. It is challenging to identify actual academic teams, nevertheless, because most definitions of academic teams that are now in use lack quantitative elements. In this study, we propose the cooperation Intensity Index (CII), a quantitative, two-way metric to assess the level of cooperation between two academics in the network for evaluating collaboration relationships. Next, in contrast to the original co-author networks, we build a new kind of co-author network with edges weighted by CII. The newest scientific research patterns, whether from university teams or elsewhere, are reflected in this network. Additionally, we suggest using TRAC (Team Recognition Algorithm based on CII) to separate academic teams from vast networks of co-authors. Lastly, we identify teams by TRAC using the DBLP data set, which comprises 1,575,949 published articles and 1,250,440 researchers. By comparing our technique with real team data and a rapid unfolding algorithm, we can show that it is successful. A vast amount of data related to scholarly endeavor, including journal articles, conference proceedings, dissertations, books, patents, presentation slides, and experimental data, are contained in Big Scholarly Data (BSD), which is rapidly expanding due to the global production of research articles by scholars as modern science advances.

In this study, Guanghong Lu, Chunhui Duan [4] et al. make the case that outlier identification is an essential component of data mining, which is widely used in a variety of domains such as fraud detection, malicious behavior monitoring, health diagnostics, etc. Global outlier identification for a collection of remote datasets is especially desired because of the enormous volume of data that is growing more dispersed than ever. In this work, we present PIF (Privacy-preserving Isolation Forest), which provides certain security assurances and has a high efficiency and accuracy in detecting outliers for many dispersed data sources. In order to do this, PIF creatively enhances the conventional forest method, allowing it to function in dispersed systems. By using a set of well-crafted algorithms, all involved parties work together to effectively construct an ensemble of isolation trees while maintaining the confidentiality of any sensitive data. Additionally, we provide a thorough architecture that can be used to both horizontally and vertically partitioned data models in order to handle complex real-world scenarios involving various types of partitioned data. We have put our technique into practice and conducted thorough trials to assess it. It is shown that PIF may preserve a linear time complexity without compromising privacy while achieving an average AUC that is equivalent to that of the current Forest. As more and more data is gathered and stored, the volume of data is essentially growing every day.

In this research, Zhao Sun [5] et al. have claimed that handling large-scale graph data is essential for a growing number of applications. Supporting fundamental graph operations like reachability, regular expression matching, subgraph matching, etc., has received a lot of attention. Graph indices are frequently used to expedite query processing. Super-linear indexing space or super-linear indexing time are often needed for the majority of indices. Unfortunately, super-linear techniques are usually never practical for very large graphs. We examine the subgraph matching issue on billion-node networks in this paper. We provide a new approach that enables effective matching of subgraphs for graphs that are stored on a distributed memory store. We handle queries using large parallel computation and fast network exploration, rather than depending on super-linear indices. Our test findings show that subgraph matching on web-scale graph data is feasible. Graphs are useful data structures for a wide range of applications, such as online, bioinformatics, and social networks. A growing number of web-scale graph applications are being used. For instance, Facebook has 800 million vertices at the moment, with an average degree of 130 for each vertex. The web network has 15 billion edges and 2.1 billion vertices by the year 2000. Large search engines are now building new infrastructure to handle a web network with one trillion vertices.

3. EXISTING SYSTEM

Finding intriguing network patterns or motifs is important and crucial in the era of social computing for a number of applications, including national security, decision intelligence, medical diagnostics, intrusion detection, social network analysis, and the identification of fake news. Sub-graph matching is still a computationally difficult task, let alone finding unique patterns within them. This is particularly true in huge, diverse real-world networks. In this study, by intelligently evaluating a user's query, we present an effective method for finding and ranking patterns of human behavior based on network motifs. Our approach makes use of the mathematical restriction that is essential for quicker detection, which is supplied by the semantics of a user's query. Our suggestion is to create query criteria according to the user's query. Specifically, we specify target patterns and their similarities using meta routes between nodes, which results in fast motif finding and ranking simultaneously. A real-world academic network is used to test the suggested approach using several node similarity metrics. The outcome of the experiment shows that our approach is resilient to the selection of similarity metrics and is capable of identifying intriguing themes.

4. PROPOSED SYSTEM

The suggested method is centered on using the RankSVM algorithm in the context of the Iris dataset to find outliers. A crucial part of data analysis is outlier identification, which finds observations that substantially differ from the rest of the dataset. We seek to illustrate the suitability of RankSVM for outlier identification tasks using the Iris dataset, which includes parameters like sepal length, sepal width, petal length, and petal width. The Iris dataset, a well-known machine learning benchmark dataset, is loaded first by the system. The data then goes through pre-processing, which includes standardization to guarantee consistency in feature scales. The RankSVM model is trained on the standardized dataset after undergoing pre-processing. By defining the problem as a binary classification challenge of inliers and outliers, RankSVM which is often employed for ranking tasks is modified here for outlier identification. Then, using the model's classification to give labels of -1 to outliers and 1 to inliers, the trained model is used to predict outliers in the Iris dataset.

4.1 Loading the Iris Dataset:

First, we load the Iris dataset, a well-known machine learning benchmark. The four characteristics of iris flowers that make up this dataset are sepal length, sepal width, petal length, and petal width. The dataset's ease of use and well defined structure make it suitable for a wide range of classification and clustering applications.

4.2 Preprocessing the Data:

Our RankSVM model must be pre-processed before it can be trained. Standardization is a typical pre-processing step in which the features are scaled to have a mean of 0 and a standard deviation of 1. By ensuring that every feature contributes equally to the model's learning process, this phase prevents any one characteristic from predominating because of scale disparities.

4.3 Training the RankSVM Model:

We next use the pre-processed dataset to train the RankSVM model. Although RankSVM is often used for ranking tasks, by rephrasing the issue as a binary classification job (inliers vs outliers), it may be used for outlier identification. The model gains the ability to differentiate between data points that are typical (inliers) and those that drastically depart from the majority (outliers) during training. The RankSVM model can efficiently detect outliers by utilizing the features of the Iris dataset, such as the connections between sepal and petal dimensions.

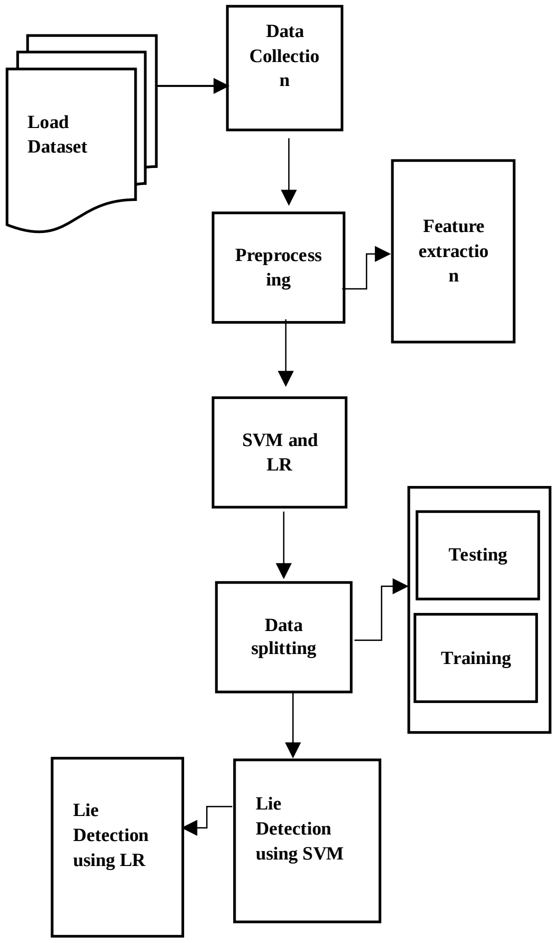

Figure 1.Block diagram

4.4 Predicting Outliers:

We employ the learned RankSVM model to forecast outliers in the dataset. The model's classification is used to identify outliers; data points labelled as -1 are considered outliers, while those labelled as 1 are called inliers. We may gain important insights into the existence of outliers and their corresponding properties within the data by using the trained model on the Iris dataset.

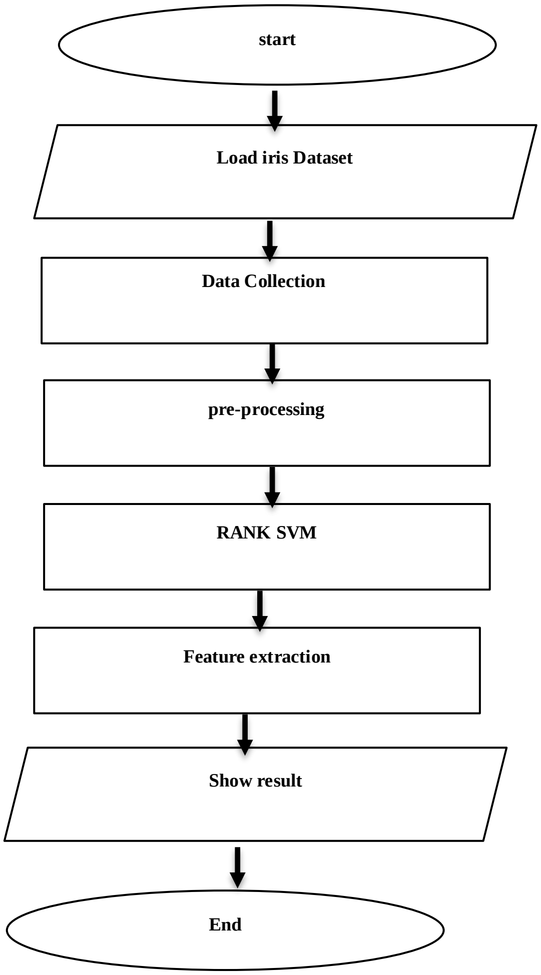

Figure 2. System Flow Diagram

6. RESULT ANALYSIS

The Iris dataset's RankSVM-based outlier identification results, when analysed, provide numerous important insights. First, the model shows its efficacy in separating outliers from data points that considerably differ from the majority by correctly identifying them within the dataset. The Iris dataset's anomalous observations' properties may be inferred from the distribution of outliers across several feature dimensions. Furthermore, quantitative assessments of the model's outlier identification accuracy are provided via performance metrics including precision, recall, and F1-score. Deeper understanding of potential data anomalies or measurement mistakes may be gained by examining the patterns and traits of the outliers that were found. Additionally, one may confirm the relevance and importance of the reported anomalies by contrasting the found outliers with expert annotations or domain knowledge. The examination of the results highlights the significance of reliable outlier detection methods such as RankSVM in revealing abnormalities and hidden insights in datasets. This enhances the quality of data and enables better decision-making across a range of applications.

Algorithm	Accuracy	Precision	Recall	F1-score
RANK SVM	0.91	0.92	0.91	0.91
LR	0.56	0.87	0.56	0.67

Table 1. Comparison table

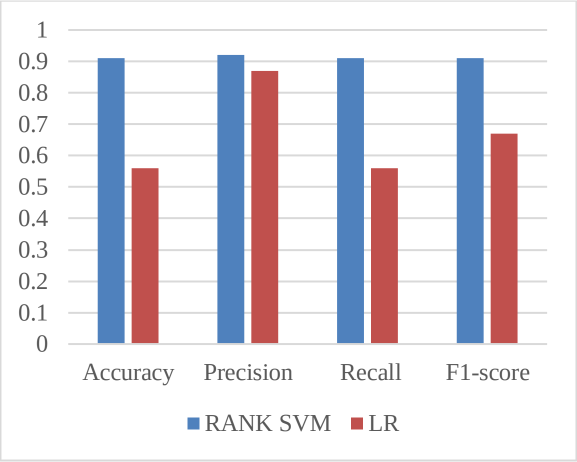

Figure 3. Comparison graph

The assessment results show the performance metrics of two outlier identification methods that were used on the

Iris dataset: Rank SVM and Logistic Regression (LR). With an accuracy of 91%, precision of 92%, recall of 91%, and F1-score of 91%, Rank SVM performs better than other models. These metrics show how well the algorithm performs at reliably locating outliers in the dataset while preserving a high degree of recall and accuracy. By comparison, LR performs far worse, with accuracy of 56%, precision of 87%, recall of 56%, and F1-score of 67%. When compared to Rank SVM, LR has a very high precision, but its lower accuracy and recall indicate that it may not be as good at catching outliers. These results highlight Rank SVM's effectiveness in outlier detection tasks and highlight its potential for use in applications that need for precise anomaly identification and classification.

7. CONCLUSION

To sum up, this research shows that the RankSVM method works well for identifying outliers in the context of the Iris dataset. Outliers that differ considerably from the rest of the sample are effectively identified by RankSVM by structuring outlier identification as a binary classification job. The approach performs robust outlier detection through pre-processing and model training, offering important insights into the existence and characteristics of outliers within the Iris dataset. RankSVM's flexibility beyond standard ranking applications is shown by its capacity to adapt to a variety of tasks, including outlier detection. In the future, more research on RankSVM and related outlier identification algorithms may improve anomaly detection techniques and deepen our understanding of complicated datasets across a range of fields.

6. FUTURE WORK

The results of this study open up various new directions for investigation in further research. First off, a more thorough grasp of RankSVM's strengths and weaknesses may be obtained by extending the assessment of its outlier identification ability over a variety of datasets outside of Iris. Furthermore, examining the incorporation of ensemble approaches or sophisticated feature engineering techniques may improve the model's capacity to identify intricate outlier patterns in high-dimensional datasets. Additionally, investigating other algorithms and strategies—like deep learning-based techniques for outlier detection—may provide fresh perspectives and perhaps boost the precision of outlier detection, especially in cases of non-linear and heterogeneous data distributions.

8. REFERENCES

[1] Weishan Zhang, Su Yang, Xinfeng Zhang, Yuan Yan Tang. 2020. A thermodynamically inspired anomaly detection tool for crowd movements in surveillance footage. Tools for Multimedia and Applications 2020), 75, 14, 8799–8826

[2] Jacob G. Grasmick, Austin Workman, Michael A. Mooney, Amanda S. Hering, and Youjiao Yu. 2018. Finding space-time anomalies in a big collection of ground deformation data. Journal of Quality Technology, 50(4), 2021, 431–445

[3] Shuo Yu, Feng Xia, Chengfei Liu, Zhaolong Ning, Jiaofei Zhong, and Kaiyuan Zhang. 2022 Team Recognition in Big Scholarly Data: Exploring Collaboration Intensity. presented at the Third IEEE International Conference on Big Data Intelligence and Computing.

[4] Jaideep Vaidya and Chris Clifton. 2020. Outlier identification while maintaining anonymity. Fourth Edition of IEEE Data Mining International Conference, pp. 233–240

[5] Bin Shao, Hongzhi Wang, Haixun Wang, Zhao Sun, and Jianzhong Li. Reliable subgraph matching for billion-node graphs. Proceedings of the VLDB Endowment 5, 9.788–799

[6] Yizhou Sun, Jiawei Han, Xifeng Yan, Tianyi Wu, and Philip S. Yu. 2020. Pathsim: Metapath-based top-k similarity search across heterogeneous information networks. Proceedings of the VLDB Endowment 4, 11, 992–1003

[7] Amy Sliva, Huan Liu, Kai Shu, Jiliang Tang, and Suhang Wang. Identification of bogus news on social media using data mining techniques. ACM SIGKDD Explorations Newsletter 19, 1, 22–36.

[8] Yizhou Sun, Jiawei Zhang, Yitong Li, Chuan Shi, and Philip S. Yu. An overview of research on heterogeneous information networks, IEEE Transactions on Knowledge and Data Engineering 29, 1 (2020), 17–37.

[9] Bita Shams and Saman Haratizadeh in 2018. trustworthy graph-based collaborative ranking. 116–132 in Information Sciences 432 (2021).

[10] Guo-Jun Qi, Thomas S. Huang, and Charu C. Aggarwal. About assembling disparate social media content by use of connections to outliers. ACM, Proceedings of the Fifth International Conference on Web Search and Data Mining, 553-562.