Comparative analysis of outlier detection approaches over cloud computing

Himanshu Rathore
Department of Computer Science and Engineering
ITM University Gwalior, India
rathorehimanshu1993@gmail.com

Arun Kumar Yadav
Department of Computer Science and Engineering
ITM University Gwalior, India
arun26977@gmail.com

***Abstract*—Outlier are those event which is drift too much after actions. So discovering outlier from a group of outline is a widespread problem in the area of data mining. The recognition of outlier can cause to locate some useful and meaningful knowledge. Previously outlier consider as noisy data, has now become frightful difficulty which has been uncovered in various domains of research. In this paper mainly focused on different kind of outlier and their detection approaches. This mainly contain classification of outlier and techniques which are classic outlier discovery method and spatial outlier discovery method. The classic outlier discovery method discover outlier in real transaction dataset, which is divided into statistical approach, distance approach, deviation approach and density approach. The spatial outlier method based on spatial datasets are separate from operation data, which are considered into spaced and graph approach. Finally, the application of outlier discovery approach.**

Keywords—Outlier mining; data mining; outlier;

I. INTRODUCTION

Data mining: The progression of mine some useful information, pattern, knowledge which is formerly unidentified and identifiable information from the huge dataset and taking it for used in organizational decision making [1]. However, there is lots of issues in mining of data is substantial dataset such as data redundancy, unavailability of data or difficult access to data, the worth of trail elements is not precise. The exist facts object that do not comply with general behavior of data which is defines as outlier. Outlier are interesting it disturbs the mechanism that causes the normal data. The numerous definition proposed for outlier, such as outlier is well-define as outline study. The identification of outlier accord beneficial, ample and significant perception and number of applications area such as climatology, ecology public health, transportation and location based services, intrusion detection, mobile phone and assurance privilege fraud recognition and industrial damage detection. Outlier may be result of changeability that is immanent in data. Recently, a few studies have been conducted on outlier detection for huge datasets [2]. Outlier techniques is very important in data filter.

The present outlier detection algorithm including based on the probabilistic and statistical, clustering, classification, depth, distance and density based methods [3]. In statistical data study of outlier dates as initial as the 19th era [4]. Since the various research groups have developed a diversity of outlier detection techniques with several of these specifically meant for certain applications. The classical outlier can be classify into statistical approach based approach, distance based approach, deviation based approach and density based approach.

II. TYPES OF OUTLIERS

A. Point Outlier

An independent fact instance can be reflected an unusual with respect to the respite of facts, then the instance is labelled as a point outlier. It is one of the easiest outlier kind and center for the many of outlier detection researcher. As an existent example, we consider credit card transaction with dataset corresponding to a single credit card transaction inferring data definition by amount exhausted. A transaction for which the amount spent is actual extraordinary related to the regular range of the regular range of disbursement for that person will be point outlier.

B. *Contextual outlier*

When a dataset is irregular through respect to certain context, then it is also as conditional outlier [5]. The belief of a context is indeed of by the structure in the dataset and has to be contemplate as a fragment of the problem formulation. Each facts instance is define using two sets of attribute:

- Contextual attributes: This is basically being is using in the direction of wind up the framework. Contextual attributes are defined as the positions, latitude and longitude in spatial dataset and in time-series data, a contextual attributes.
- Behavioral attributes: This is well define as the quantity of rainfall at a particular place, when stated about the middling rainfall of the whole planet within the domain of geospatial data.

C. *Collective outlier*

When related datasets collection is compare to the entire instance, than it is well-defined as a collection outlier.

III. VARIOUS OUTLIER DETECTION

Outlier detection has been substantial studied in the past decennium and innumerable method have been produce. Outlier approach is distinguish in two group: Classic outlier method analysis outlier based on operation dataset and spatial outlier method analysis outlier based on spatial dataset.

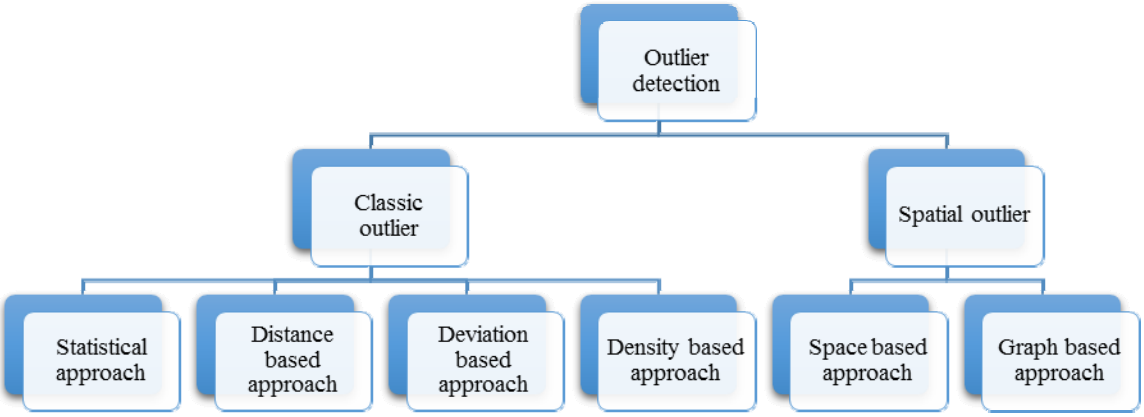

Fig. 1. Classification of outlier approaches

A. *Classic Outlier*

This type of outlier based on operational dataset which is collection of numbers of items. A typical case is "market basket data", in which every transaction is considered as the cluster of items asset by the user in operation. Such information can also be amplified by auxiliary item relate the customer or the context of transaction. As a result, mostly outlier are used for research on trading data.

- Statistical based approach: The oldest procedure which have been used for outlier identification, which a model coined as distribution model for give dataset and using a divergent test they detect outlier. Many techniques are being account in both Branett ad Lewis [6]. Rousseuuw and Leroy [7] are solitary dimension. Reusseuuw expecting dimension it become more complex and imprecise to make a model for dataset.
- Distance based approach: Regular data items have more dense neighborhood, outlier are for away each other i.e. has a less dense neighborhood and has been 1st coined by Knorr and Ng [8]. Distance based approach is not produce entail knowledge about a ranking of outlier detection but it is used to define a preferable rank of the parameter.

- Deviation based approach: In deviation based method [9] given a set of information point outlier are point that do not fit to the general feature of that set. This method is used in chronological problem where outlier are acknowledge by the uses of tropical data point quality and deviated data quality. Jagadish et al. give an approach now relation to histogram.
- Density based approach: This approach evaluate the density distribution of the data and find the outlier which are in low-density territory. Brewing et al. [10] allocate a Local Outlier Factor (LOF) to for each point passed on the native compactness of its neighboring density of domain, which is resolve by user minimal numbers of point (MinPts). Papadimitriou et al. [11] current Local Correlation Integral (LOCI) which usages numerical standards based on the information itself to handle issues of preferring values for MinPts.

B. Spatial Outlier

Classic approach have to be remodel due to the qualitative difference between spatial and non-spatial feature:

Spatial dataset could be labeled as a reference objects of cluster of spatially. On the based on spatial feature it two group one include the location, shape and other topological feature other contains length, owner, building, age and name. The acknowledgement of spatial outlier disclose concealed but precious information in many application.

- Space based approach: This approach apply Euclidian distance for selecting spatial neighborhoods. Yufeng Kon et al. profound spatial weighted outlier detection algorithm which use feature such as cluster distance and mutual border length as weight when associating non-spatial attributes [12] and another algorithm is been profound by Adam et al. covering spatial as well as semantic association amongst neighbor [13]. Lieu et al. profound a method for finding outlier in unusual distributed spatial dataset [14].
- Graph based approach: This approach is mostly used to related graph connectivity to describe spatial neighborhoods. Yefung kon et al. profound a set of graph based algorithm to discover spatial outlier, which foremost create a graph based on k-nearest adjacent association in spatial domain, and continually cut upraise mass boundaries to recognize in accessible point or edges which are highly divergent to their adjacent item.
- Cluster based approach: This outlier detection approach is relatively effective as the data from the datasets is initially segmented into clusters. In every cluster each data point is authorized as a degree of the membership. The outlier is discovered without any interference in the clustering process. Numerous clustering approaches are used for the outlier detection [15]. Clustering on streaming data is classified by grid based and k means/k median methods [16]. In Partitioning methods, various centroid based approaches, k means, PAM (Partitioning Around Medoids), CLARA (Clustering LARge Applications) and CLARANS (Clustering Large Applications based on RANdomized Search) etc. methods are used [17]. One of the clustering is hierarchical clustering. In it, the entire data set is further decomposed into dissimilar small datasets. It is further divided into two categories i.e., Agglomerative methods (in which sample units are united to form single cluster) and divisive methods (in which single parent cluster is further partitioned) [18]. An Agglomerative method generally starts with each point as a distinct cluster and it association two closest clusters in each succeeding step until the verge condition is met. A divisive method, contrary to an agglomerative method, begins with all the points as a single cluster and splits it in the next succeeding step until the threshold condition is met. Agglomerative methods are more popular in use. [19]. Various Hierarchical Methods are MST clustering, CURE and CHAMALEON. Moreover, in very large databases, BIRCH [20], is used and it can be enhanced for higher dimensional data.

This algorithm have major advantage compared with the existing spatial outlier methods: exact in detecting point outlier and ability to identifying territory outlier [14].

IV. APPLICTION OF OUTLIER DETECTION

A. Fraud Detection

An application here would attract to maintain a custom dataset of every client and check the database to identify any for changes describe as monitoring of motion. In outlier techniques are useful for detection in credit card duplicitous application and credit card duplicitous usage.

B. Insider Fraud Detection

Within data may be unusual types, generally denotes to any of the data that can influence the cost of stock in a given manufacturing pending legislation affecting an exacting industry.

C. *Intrusion Detection*

Outlier detection techniques help into find out the suspicious different types of transaction in government department i.e. IT department. It is also describe recognition of odious motion in the system model to computer [21] motivating from a computer protection outlook.

D. *Industrial Damage Detection*

Industrial unit suffer damage due to regular usual wear and tear. This kind of damages need to be detected early to prevent further rise and losses. The information in this field is generally sensor recorded data by means of various sensor and composed for investigation.

E. *Therapeutic and Community Health Outlier Detection*

In such scenarios, for identifying outlier the method used in grouped outlier detection. Numerous methods also focused on identifying diseases in a definite locale [22].

V. DIFFICULIES IN OUTLIER DETECTION

- In numerous zones regular activity keeps growing and might not be present to be an agent further.
- Conflicting concept of outlier in diverse applicable make it complex to relate method formed in a single arena to the other.
- Imprecise margin between common in addition to outlier performance at a deviation of outlier examination line near to the edge could really be regular and vice-versa.
- Approachability of labeled information for instruction of models which are being worn by outlier detection methods.
- Honk in the information which tends in relation to the outlier of original nature, this is why it complicated in differentiate and eliminate.

VI. RELATED WORKS

Howkins [5] focused on, The classic explanation if an outlier is due to who express an outlier is an observation which diverges so much from other observation as to provoke uncertainties that it was engendered by a diverse mechanism.
Hawkins defines outlier study is very different respect to other available study, which it looks like it was gained through a different method. Lazarevic & Kumar anticipated a restricted algorithm by outlier detection with a practice known as feature bagging. Shekhar et al. anticipated the explanation: A spatial outlier is spatially related things whereof non spatial rank values are most distinct to those of other spatially assign recipient with their spatial neighborhood. A varies survey, review articles and text books particularly Hodge & Austin, wrap outlier detection methods in in different domains. Numerical and figurative data approaches, numerical approaches have been presented by different researchers. Cyber-intrusion recognition survey and research and evaluation books on methods by outlier detection are exceptional ways of text on the topic.

Nishant Gaur [23] introduced an outlier discovery generally focuses on particular area of research domain or on a self-application. Initially, we reviewed several work related to outlier detection and there after we made comparisons to discuss various applications and techniques of it. Distinction in simple also complex outliers and then defined types of complex outliers.

Knorr et al. [26] proposed a new explanation based on the conception of distance, which regards a point p in dataset as an outlier with respect to the parameter k and l (lambda), if no more than k point in the dataset are at a distance l (lambda) or less than p. A large extent of survey, review articles and test books particularly Hodge and Austin [6]. Wrap outlier detection process in the different domains. Numeral and figurative data approaches have been accessible by different researchers.

R. Subramanium et al. [25] introduced about the clustering algorithms, which are K-means and Expectation-Maximization algorithms. In K-means was not assurance convergence while Expectation-Maximization's quick convergence. Their experimental results shows that EM approach has faster execution time than k-Means approach, by using linear regression with an explanatory variables. Regression models also describe the defference in the dependent variable. And the result is stored into voronoi cell, the efficient heuristic algorithm are generally used to get local optimum equal to the E-M algorithm for combination Gausian distribution by iterative refinement techniques deployed at both algorithm which is K-mean. K-mean which has introduced context-sensitive clustering techniques depend on the bayes decision theory to measure on unsupervised way to

statistical parameter of the class nearly used in the Bayesian decicsion rule. It improve result of EM clustering with compared to K-mean techniques. Currently model based algorithm are used to be set right efficiency of clustering algorithm. Graubs statistics defines an outlier in a collection of data as an observation for subset which catches be incompatible with the remainder of that set of data.

Maitreyee et al. [24] proposed that the used of partitioned algorithms with distance based approach to outlier detection has advantage of adding algorithm the unsupervised method which has the property of testing the outlier in future with an efficient manner whenever the new data will be added to database, which increases time involution of the detection process. It say that the CLARANS and FCM process are superior as compared with PAM and CLARA. On the basses of performance, when PAM and CLARA both are same. Thus, it could be deduced so the CLARANS algorithm is accurately detecting the outliers. The time complexity show in order to cost and size of the dataset, have less computation cost, as only some iteration are required and makes PAM an impractical solution for large datasets. It can be say that fastest algorithm is CLARANS, followed by CLARA and PAM. The execution of CLARA and PAM is very close.

S.D pachgade et al. anticipated an efficient outlier, outlier detection method in which they firstly group the dataset into numbers of clusters. It will decreased the computational time, then by using threshold value from user and calculated outlier according to given threshold rate for each cluster. This techniques has limitation, it alone thing with the numerical data not with textual mining.

Kamalijeet Kaur correlate the execution of cluster techniques, in which all comparative result is depends on the number of the dataset. Clustering techniques has high performance and low computational complexity. The clustering approach of outlier detection for high structural data which is helped in discovery of entity which are dissimilar, distinctive and mismatched with present data. To increase the dimensionality of the dataset connecting to the nearest neighbor, the concept of hub was developed. Clustering plays an important role in handling the high performance data and is one of the techniques of the outlier detection. Computation time which decreased by cluster approaches are used by author are k-means and fuzzy c means for finding out the outlier. The anti-hub point is embedded in the clusters that are formed after using this clustering method. It is concluded that the anti-hub that is applied into K–Means is more efficient than the anti-hub applied to Fuzzy C Means. Proposed a clustering based technique which used K-Mean clustering algorithm for clustering of the datasets and density based and distance based algorithms for consequence out from the outlier and also in case of bup a dataset experimental analysis of lower and higher dimensional data has outlier clustering algorithm can be used for outlier analysis.

Vijay defines algorithm for detection of outlier using cluster based approached, in this approach first they do PAM clustering algorithm. After that small clusters are determined and consider as an outlier cluster. Then using absolute distance between the Mediods of the current cluster and each are of the point in the cluster, through this calculation detecting the rest of the outlier (if any) [17].

PAM clustering techniques has been used for clustering the data, it contain datasets having two dimensions, then the PAM s analyzed on the high dimensional data, so here another take Bupa data set. Initially started with k-means with OFT which gives more outlier detection in case of higher dimension data that in outlier detection using PAM. K-means has sensitivity over outlier data but can be still used OFT for the detection of outlier data [15].

VII. Comparitive Analysis

As many of outlier detection approaches are available for detecting an outliers and the usage of all these vary according to which kind of data is being used, size of the dataset etc. [27]. Now with the remarkable progress of data, best outlier detection approach have to be applied on large data sets [28]. So the elementary parameters like efficiency, computational cost, scalability and applicability needs to be studied.

A. Efficiency

It is well-defined as the estimation of the average execution time needed for an approach to complete work on a particular data set. Efficiency of a process is measured by its order. It is helpful for quantifying implementation difficulties of definite problems.

B. Computational Complexity

It is directly comparative to the computational complexity of the approach. It is the estimation of the number of steps required by the approach associated for input of an instance or a given size in the worst case. Functional size is estimated by the various steps.

C. *Scalability*

It is defined as the ability of the product or a computer application to continue to function well even when it is altered in size or volume, according to the user needs. It is basically a rescaling like expandability of an application platform which can be used on larger operating systems for handling large number of users and also for better performance.

D. *Applicability*

As each approach has its boundaries and limits set for being applicable on any given set of data. Depending upon the data set i.e. whether it is a statistical data or large dataset, various approaches are applied on the datasets to detect outliers. All the above stated outlier detection approaches are compared in table-1 with respect to certain parameters like efficiency, computational cost, scalability, applicability and high dimensional data etc.

TABLE 1. Comparison of outlier detection approaches

	Efficiency	**Scalability**	**Computational Complexity**	**High Dimensional data**	**Feature spaces**	**Applicability**
Statistical based approach	2	1	5	0	Univariate	Statistical Data
Depth based approach	2	1	5	0	Univariate	Statistical Data
Distance based approach	3	2	4	2	Multivariate	Based on the distance points
Density based approach	4	4	5	2	Multivariate	Based on the neighborhood of the data point
Clustering based approach	5	4	2	3	Univariate / Multivariate	Based on cluster of data

Note: 0- Not applicable, 1- very less, 2- less, 3- average, 4- high, 5- very high

VIII. RESULT COMPARISION

The above graphs in fig.2-5 is designed by evaluating, analyzed and compared in the table1 which show the variation in terms of efficiency, high dimensional data, scalability, feature space, computational complexity and applicability for all outlier detection approaches on huge datasets.

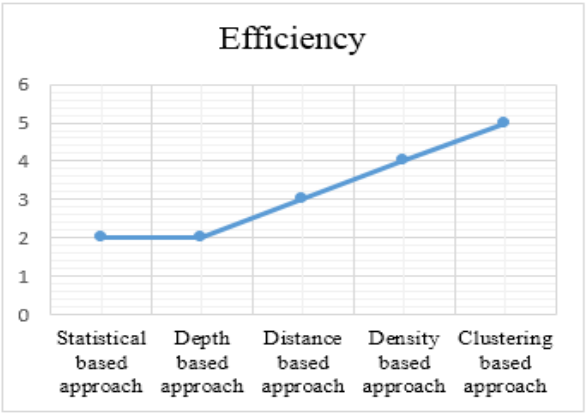

Fig. 2 Comparison of efficiency for outlier detection approaches

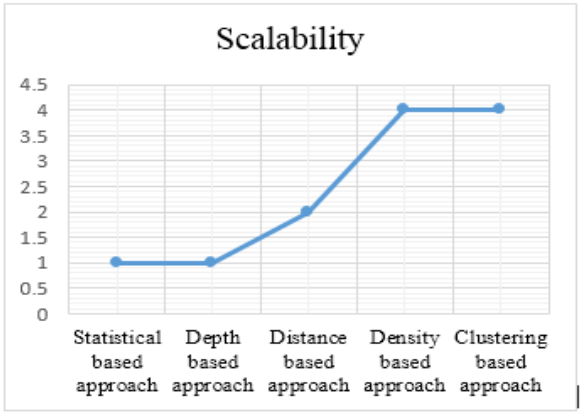

Fig. 3 Comparison of scalability for outlier detection approaches

Himanshu Rathore et al. / International Journal on Computer Science and Engineering (IJCSE)

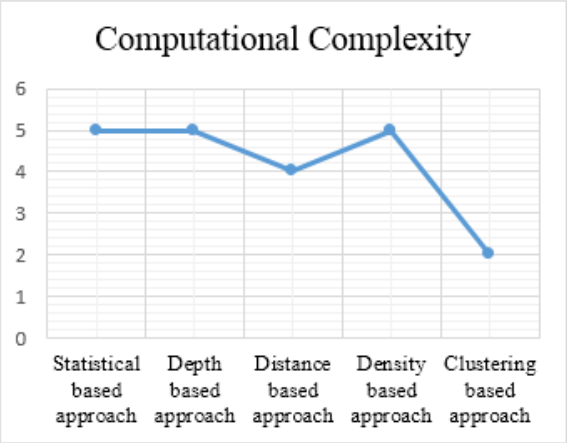

Fig. 4 Comparison of complexity for outlier detection approaches

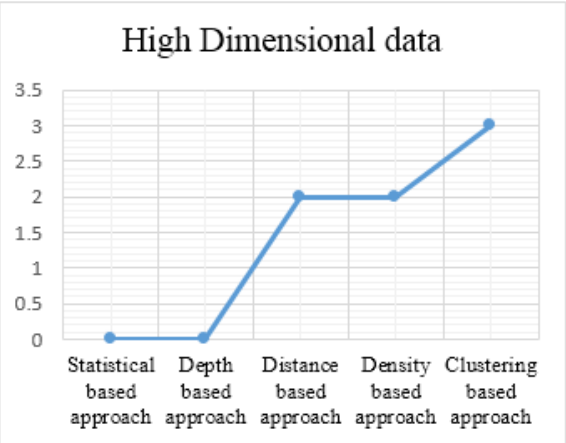

Fig. 5 Comparison of data for outlier detection approaches

In fig. 2, efficiency statistical outlier detection approach and depth outlier detection approach are less efficient with respect of other three approaches in which clustering outlier approach is highly efficient outlier detection approach. From fig. 3, The comparative graph show that the statistical outlier detection approach, depth outlier detection approach and distance outlier detection approach are less scalable with compare to other two approach which is density outlier detection approach and clustering outlier detection approach which are high in scalability. And the last parameter i.e. computational complexity, clustering outlier approach is best approach because it has low complexity, highly scalable and highly efficient which is shows in fig. 4. And in terms of high dimensional data the fig. 5 reflected that statistical based approach and depth based approach have very low dimensionality in term of data whereas clustering based approach has high dimensionality in terms of data. The comparative analysis reflect that clustering outlier detection approach is optimized for large datasets.

IX. Conclusion

The problem of outlier detection finds applications in numerous domains, where it is desirable to determine interesting and unusual events in the underlying generating process. The core of all outlier detection methods is the creation of a probabilistic, statistical or algorithmic model that characterizes the normal data. The deviation from this model are used to identify the outlier. A good domain specific knowledge of underlying data is often crucial designing simple and accurate models that do not over fit the underlying data.

In this paper, I have bring collectively a diversity of techniques in an entirely idea and extensive illustration about outlier detection approach from data mining aspect. Then I portrayed about the different types of outlier and their identification which can be classified based on two categories: - classic outlier approach and spatial outlier approach. And also includes application and difficulties or problem in outlier and outlier detection. Outlier analysis has tremendous scope for further research especially in the area of structural and temporal analysis.

References

[1] Han and Kamber, data mining: Introduction to data mining, Moghan Kaufmann publisher, 2nd edition, PP. 5-6, 2006.
[2] Lazaraic, A. Kumar, feature bagging for outlier detection in KDD, 2005.
[3] Peigue fuand and Xiaohni Hu. "Biased sampling of density- based local outlier detection algorithm", *12th International conference on natural computation, fuzzy system and knowledge discovery*, 2016.
[4] Edgeworth F.Y., on discordant observations philosophical magazine, PP.5, 23,364-375, 1887.
[5] D. M Hawkins, "identification of outlier", Chapman and hall Landon 1980.
[6] Noble, C.C and Cook, D.I., "Graph based outlier detection" In preceding *9th ACM SIGKDD International conference on knowledge discovering and data mining*, ACM Press, 631-636, 2003.
[7] Barnett, V. and Lewis, T, "Outlier in statistical data", John Willey and Sons, 3rd edition, 1994.
[8] Rousseeuw and Leroy, "Robust Regression and outlier detection", John Willeyand Sons, 3rd edition 1996.
[9] E. Knorr, R.Ng, and V. Tucakov, "Distance-based outlier. Algorithm and application", *VLDBJ*, Vol.8, nos.3-4.2000, PP. 237-253.
[10] Brewing M.M., Kriegel, H. P., and Ng R.T.,"LOF: Identifying density based local outlier." *ACM Conference Proceedings*, 2000, pp.93-104.
[11] Papadimitriou, S., Kitawagh, H.,Gibbons, P. ,Falouslcos,C., "LOCI: Fast outlier detection using the local correlation integral", in Proceeding *international conference on data engineering*,2003.
[12] Y. kuu, C,-T.lu, and D.Chen. "Spatial weighted outlier detection" in preceding *6th SIAM international conference on data mining*, Bethesda, USA, 2006, pp.614-618.
[13] N.R Adam, V.P Janeja, and V.Atteri, "Neighboorhood based detection of anomalies in high-dimensional. P Spatiotemporal sensor datasets."
[14] H. Liu, K. C Jezek, and M. E O'Kelly,"detecting outlier in irregularly distributed spatial dataset by locally adaptive and robust statistical analysis and GIs", *International Journal of geographical information science*, vol. 95 issue 8, pp. 721-741, 2001.
[15] H. S. Behera, "A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 4, pp. 287-292, April 2012.

Himanshu Rathore et al. / International Journal on Computer Science and Engineering (IJCSE)

[16] F. Anguiulli, F. Fassetti, "Detecting Distance-Based Outliers in Streams of Data", In Proceedings of the *16th ACM Conference on information and knowledge management (CIKM)*, 2007 , PP. 811 – 820.

[17] Vijay Kumar, "Outlier Detection: A Clustering-Based Approach", *International Journal of Science and Modern Engineering (IJISME)*, Volume-1, Issue-7, pp 16-19, June 2013.

[18] W. Fan, M. Miller, S. Stolfo, W. Lee, P. Chan "Using artificial anomalies to detect unknown and known network intrusions", In Proceedings *IEEE International Conference on Data Mining, IEEE Computer Society*, Volume 6, Issue 5, April 2004, pp. 507-527.

[19] Y-Shi, "COID: A cluster- outlier iterative detection approach to multi-dimensional data analysis", *Knowledge Information System*, Volume 288, No 3, pp. 709 – 733, 2011.

[20] T. Zhang, R. Ramakrishnan, and M. Livny. "BIRCH: An Efficient Data Clustering Method for Very Large Databases". In proceedings *ACM International Conference on Management of Data (SIGMOD'96)*, Montreal, Canada, pp. 103-114, 1996.

[21] Yufeng Kou, Chang-Tien lu, Dos santos, R.F, "Spatial outlier detection: A graph-based approach". *ICTAI*, Volume 1, pp.281-288, 2007.

[22] Keogh and Herle. *18th IEEE sypnosium*, USA, PP.329-334, 2005.

[23] Nishant Gaur and Rashi Bansal. "Outlier detection: application and techniques in data mining," *International Conference- Cloud System and Big Data Engineering (Confluence),* 2016, pp. 373-377.

[24] Poonam and Maitreyee Dutta. "Performance analysis of clustering methods for outlier detection," *Second International Conference on Advanced Computing & Communication Technologies,* 2012, pp. 89-95.

[25] R. Subramanian and V. Vasudevan. "Outlier detection for regression using k-means and expected maximization methods in time series data," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3 issue 9, pp. 1052-1059, September 2013.

[26] Dhaval R. Chandaran and Maulik V. Dhamecha. "A survey for different approaches of outlier detection in data mining," *International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO),* 2015.

[27] Rajendra Pamula, Jatindra Kumar Deka, Sukumar Nandi, "An Outlier Detection Method based on Clustering", *Second International Conference on Emerging Applications of information Technology*, 2011, pp. 253-256.

[28] D. Xiang, W. Lee, "Information-theoretic measures for anomaly detection", In proceedings of *IEEE Symposium on Security and Privacy*, pp. 130-143, May 2001.

AUTHORS PROFILE

Himanshu Rathore has completed B. Tech. in Computer Science and Engineering. He is currently pursing his M. Tech. in Computer Science and Engineering from ITM University Gwalior, Madhya Pradesh, India. His research of interest includes cloud computing and Data Mining.

Arun Kumar Yadav has done B.E. (Computer Science & Engineering) from G.B. Pant Engineering College, Pauri Garhwal, M.Tech (Information Technology) from Sam Higginbotom Institute of Agriculture, Technology & Sciences, Allahabad and pursuing Ph.D. in Computer Science and Engineering from Uttarakhand Technical University, Dehradun. Presently he is working as Associate Professor in the Department of Computer Science & Engineering at ITM University Gwalior, Madhya Pradesh, India. His research interest includes Distributed Database Security, Cloud Computing and Fog Computing. He is a senior member of IACSIT and IAENG Technical Societies.