Acta Scientiarum
http://periodicos.uem.br/ojs
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v46i1.66902

The influence of feature grouping algorithm in outlier detection with categorical data

Sharon Femi Paul Sunder Nathaniel*, Kala Alwarsamy, Rajalakshmi Viswanathan, Ganesh Vaidyanathan Subramanian and Vidhya Veerabahu

Sri Venkateswara College of Engineering, Pennalur, Sriperumbudur- 602117, India. *Author for Correspondence: E-mail: sharon@svce.ac.in

ABSTRACT. Outlier mining has become a rapidly developing domain over the recent years with increasing importance in the fields like banking, sensor networks, and health care. In general, anomaly detection methods are compatible with numerical data and ignore categorical data. However, in real-time problems, both numerical and categorical data are to be considered to obtain accurate results. There are several methods available for the outlier detection of high dimensional data in numerical data. In this paper, a feature grouping algorithm for anomaly detection is proposed that considers the categorical data also. This algorithm correlates the features of categorical data and forms feature clusters and detects the outliers. The features are assigned feature weights based on their levels of appearance and the outlier scores are determined. The performance of the feature grouping algorithm is then compared with the traditional algorithms like LOF and Isolation Forest algorithm and state-of-the-art methods like WATCH on UCI datasets. From the experimental evaluation of the results obtained, it is found that the proposed algorithm is comparatively better than the existing algorithms for categorical data.

Keywords: outlier detection; feature grouping; categorical data; lof; isolation forest.

Received on January 30, 2023.
Accepted on September 12 2023.

Introduction

Outlier detection is the identification of abnormalities that cause catastrophes if it is left undetected. The standard definition for outliers given by (Hawkins, 1980) says that it is an observation that deviates so much from other observations as to arouse suspicion, when compared to the results generated by other different mechanism. Outliers are of two types. In the first type, outliers are considered as noise that has to be eliminated during pre-processing of data. In the second type, outliers are the target that has to be determined. This is primarily important in data mining and has to be detected so as to avoid catastrophic effects. The main applications of outlier detection include detecting frauds in financial transactions, detecting diseases in healthcare, intrusion detection, spam filtering and so on (Wang, Bah, & Hammad, 2019).

Numerous research is carried out to examine the concepts, approaches, difficulties, results and potential directions of the current outlier detection algorithms (Chandola, Banerjee, & Kumar, 2009; Femi & Vaidyanathan, 2018; Hodge & Austin, 2004). There are many existing algorithms for finding outliers, which are categorized as statistical based, distance based (Knorr & Ng, 1998), density based (Tang & He, 2017) and clustering based methods (Shi & Zhang, 2011; Wang, Jiong, Su, & Qian, 2019). In one of the cases, the sequential and parallel models are combined to form an ensemble framework in detecting the outliers (Femi & Vaidyanathan, 2022). Though many of the algorithms are available for outlier detection, yet there may be limitations in implementing them only for numerical attributes by ignoring the categorical attributes. However, very few of the algorithms in the literature handle categorical data (Eiras-Franco, Martínez-Rego, Guijarro-Berdiñas, Alonso-Betanzos, Bahamonde, 2019; Hu, Wang, Cheng, 2019; Li, Zhang, Qin, Xun, 2019), but they also suffer from low detection precision and high time complexity.

The existing outlier detection system detects outliers only to remove the abnormal events that cause complications while training a machine learning model. The existing systems also focus only on the outlier detection of continuous attributes or numerical attributes in which the categorical data are either ignored or are converted to some equivalent numerical value. This may lead in loss of data or neglecting the significant features of obtaining the accuracy. Therefore, categorical data are mapped into numerical value to avoid loss

of data. In addition to that, the data in majority of the practical applications are in categorical form. Thus, it becomes necessary to detect outliers for categorical data for effective usage.

Based on the above highlights and considerations, the proposed work is intended to study the influence of outlier detection of categorical data by feature grouping algorithm. The features of the data instances are considered to be the feature set. The correlation between the features is determined and the similar features are clustered. Feature weights are then assigned and the outlier score is evaluated. Candidate set of outliers are obtained and various metrics are evaluated and the results are compared with LOF and Isolation Forest. The novelty of the proposed method is that correlation between all the features is examined and closely related features are grouped into a various feature groups.

The main contributions of this paper include,

- Determining the correlation between the features for feature grouping
- Determining the outliers in categorical data
- Comparing the performance of the proposed method with traditional algorithms and the state-of-the-art algorithms.

The remaining part of the work is described as Section 2 analysis of literature review of the work, section 3 details the proposed method, section 4 discussions of result and the comparison of performance of algorithm and section 5 present the conclusion of the work.

Related works

This section deals with the outlier detection methods in the literature handling categorical attributes. Detecting outliers in qualitative variables has fewer research solutions compared to quantitative datasets.

There are a few methods that have been used to identify outliers in high-dimensional data. The contextual outliers are explored in relevant subspaces and outliers are discovered in subspaces rather than in multi space (Zhang et al., 2016; Zhang, Yu., Xun, Zhang, & Qin, 2017). The outlier scores calculated by the various outlier detection algorithms are then merged to locate the quality outliers using the feature bagging method (Lazarevic & Kumar, 2005), which combines the findings from many outlier detection algorithms. Outliers are investigated from a novel angle using a multi-view low-rank analysis (MLRA) framework for outlier detection from multi-view data (Li, Shao, & Fu, 2018). In addition, the majority of high-dimensional datasets combine information from several sources of measurements and observations. The extensive feature group information can only be partially utilized by current algorithms for high-dimensional data to identify outliers.

Anomalies in categorical data are detected by a method called Neural Probabilistic Outlier detection (NPOD) (Cheng, Wang, & Ma, 2019). The log-bilinear neural model is employed that captures the relationship between the categorical attributes. The learning loss of the neural network model separates the inliers from the outliers. In this, two indicators are used in determining the outlier score.

Two novel outlier detection algorithms are developed for detecting outliers in datasets with categorical attributes (Du, Ye, Sun, Liu, & Xu, 2020). These are based on entropy difference threshold which employed heuristic strategy for selecting the threshold. The first algorithm is designed for small datasets. It constructed Outlier Detection Tree (ODT) that classifies every data point as normal or abnormal class using if then rules. The second algorithm is the advanced version of ODT, called FAST-ODT which outperformed the existing methods in terms of detection accuracy and time complexity.

Three approaches are employed for ranking unsupervised outliers by considering categorical and numerical attributes together (Garchery & Granitzer, 2018). Among them, two approaches are entropy-based methods depending on individual and collective entropy and the third approach extends the Isolation Forest to support mixed data. These entropy-based methods spotted more global outliers than local outliers. The outliers identified in the mixed space are different from the outliers obtained by considering only numeric attributes. The results present the influence of including categorical attributes in ranking

A semi supervised novel approach is introduced for anomaly detection on categorical data (Ienco, Pensa, & Meo, 2016). This either depends on the supplementary information about the data or based on the type of data to be manipulated. By inferring the manner in which the two values of a categorical attribute co-occur, the distance between them is calculated. It uses a distance-based method to differentiate the anomalous instances from the normal instances of the data.

A parallel outlier mining method called POS is developed, for high dimensional data considering the categorical attributes (Li, Zhang, Pang, & Qin, 2018). It is based on feature grouping and comprised of two modules, one for parallel feature grouping and the other for parallel outlier mining. POS was implemented in spark environment and found to provide high performance.

An approach of relative patterns is proposed for enhancing the outlier detection in categorical data (Pai, Wu, & Hsueh, 2014). When frequent item set mining is used, it suffers from distortion. To overcome distortion, the relative patterns discovery approach is employed which involves hash-index based intersecting approach (HA) and an unsupervised approach (UA). Normally, UA helps to determine the anomalous observations using the knowledge of relative patterns.

Three different categorical data clustering algorithms: K-modes, STIRR, and ROCK are analyzed to detect outliers based on various parameters (Nowak-Brzezińska, & Łazarz, 2021). They conducted experiments on datasets with varying numbers of objects, variables, and categories of qualitative variables to evaluate the performance of the algorithms. They aimed to determine if the algorithms consistently detect outliers and how much they depend on individual parameters and dataset characteristics.

A comprehensive survey is conducted on outlier explanations, bridging gaps in the current literature concerning outlier detection (Panjei, Gruenwald, Leal, Nguyen, & Silvia, 2022). The challenges of producing outlier explanations and the ground truth in evaluating them are explored and the existing techniques that tackle these challenges are examined. They categorized the outlier explanations into three classifications: the significance levels of outliers, the causal interplays among outliers, and the attributes that stand out. They also elaborated on the significance of each type, how they connect to non-evaluative aspects of explanations, applications of outlier explanations and the methodologies employed to evaluate them.

An approach for estimating density to detect and clarify abnormal values within categorical datasets is introduced (Angiulli, Fassetti, Palopoli, & Serrao, 2022). This method employs measures like frequency occurrence and cumulated frequency distribution to identify outliers, encompassing both lower outliers (remarkably infrequent values) and upper outliers (extraordinarily frequent values) and provides interpretable explanations for the detected anomalies.

An outlier detection algorithm is proposed for categorical matrix-object data, which addresses the problem of outlier detection in datasets where objects are described by multiple feature vectors (Cao, Wu, Yu, & Liang, 2021). The algorithm defines the coupling and cohesion of matrix-objects based on distance, information entropy, and mutual information, and calculates the outlier factor of each matrix-object. Experimental results on real and synthetic datasets demonstrate the effectiveness of the proposed algorithm in detecting outliers in matrix-object datasets compared to other algorithms.

From the detailed literature review, it is understood that the existing outlier detection algorithms may not be well established for outlier detection in categorical data. Therefore, in the present work, the feature grouping method is employed for detecting the outliers effectively in categorical data and the obtained results are compared with LOF, Isolation Forest and state-of-the-art methods

Proposed method

In the proposed method, the attributes or features in a data set are grouped into clusters based on similarities and differences. Those instances with features that do not fit into the clusters are classified as outliers. The features are grouped and all the features are assigned feature weights based on their characteristics. The feature weights play an important role in the outlier detection. The feature weights are also utilized for calculating the outlier scores. The algorithm is evaluated for its performance by comparing with the traditional algorithms namely LOF and Isolation Forest and the state-of-the-art algorithm, WATCH. The proposed architecture diagram is shown in Figure 1.

LOF based outlier **detection**

LOF is a density-based clustering method for detecting local outliers (Breunig, Kriegel, Ng, & Sander, 2000). This approach measures the outliers of the dataset with the aid of associating an outlier score referred as the Local Outlier Factor (LOF) to every object within the dataset. LOF is characterized by MinPts, a parameter that indicates the neighborhood of every data item that is considered.

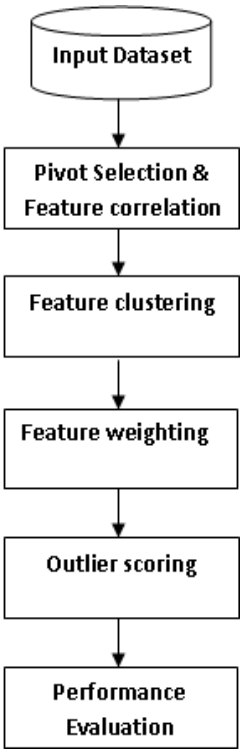


Figure 1. Proposed outlier detection model.

Let $d(x_j, k)$ be the distance of x_j to its kth nearest neighbor in $N(x_i, k)$ and $Dist(x_i, x_j)$ be the distance from x_i to x_j. The reachability distance is the maximum distance between $d(x_j, k)$ and $Dist(x_i, x_j)$ and expressed as $reachdist_k(x_i, x_j)$. This is defined as,

$$reachdist_k(x_i, x_j) = \max(d(x_j, k), Dist(x_i, x_j)) \quad (1)$$

The average reachability distance between x_i and all other data points x_j in its neighborhood $N(x_i, k)$ is defined as,

$$avgreach(x_i) = \frac{\sum_{x_j \in N(x_i,k)} reachdist_k(x_i, x_j)}{k} \quad (2)$$

The local reachability density is defined as the inverse of the average reachability distance and is given by,

$$lrd_k(x_i) = \frac{1}{avgreach(x_i)} = \frac{k}{\sum_{x_j \in N(x_i,k)} reachdist_k(x_i, x_j)} \quad (3)$$

The LOF score is calculated using the local reachability density and is defined as the ratio between the average lrd of its neighborhood and the lrd of the data point.

$$LOF_k(x_j) = \frac{\sum_{x_j \in N(x_i,k)} lrd_k(x_j)}{lrd_k(x_i) * k} \quad (4)$$

In this technique, a high value of LOF score for a data point indicates that the data point has deviating density compared to its neighborhood indicating that the data point is an outlier.

Isolation forest

Isolation Forest is based on the property of anomaly in which the anomalies are "few and different" (Liu, Ting, & Zhou, 2008). The algorithm isolates the anomalies from the normal data points and constructs k-number of 'iTrees', where k is taken as input. Since the anomalies are few and different, they occur close to the root node. The nodes with small average path length values are considered as anomalies. This procedure involves two phases. In the first phase, iTrees are constructed using the training set and in the second phase, the test objects are sent to the iTrees to obtain the outlier score of every object. Finally, the objects with score close to 1 are considered as outliers. The input categorical data is encoded using Label Encoding to covert the categorical data to numerical data. The output of this Encoding module is passed through the Isolation Forest module, where the outliers are detected.

Proposed feature grouping method

Given a categorical dataset CDS, let X = {x_1,x_2, ...x_n} be the set of n objects of CDS and Y=={y_1,y_2, ...y_m} be the set of m categorical features. The aim of the feature grouping method is to place each of the m categorical features into one of the r feature clusters, C = {C_1, C_2, ...C_r}, where every feature cluster, C_i, holds highly correlated features and are distinct from one another. Once the feature clusters are formed, feature weights are calculated and assigned based on their importance. These feature weights are further used to compute the outlier scores from the list of identified outliers.

Feature correlation

Feature correlation determines the features that are closely related that represent the dense region and by determining the dense region, the unwanted features can be pruned and the sparse subspace is searched to determine the local outliers (Femi, Vaidyanathan, & Kala, 2021). In the proposed method, feature groups are formed by identifying the similarities and differences among the features. The correlation between features is found using Cramers'V coefficient. This coefficient depends on the nominal variation of Pearson's Chi-square Test. It is used to measures the association between two categorical features and is given by Equation 5.

$$V = \frac{\frac{X^2}{N}}{MIN((p-1,r-1)} \quad (5)$$

Where, X is derived from chi-squared test, N is the total count of observations, p refers to the number of columns or features and r represent the number of rows. The range of the output is [0,1], where 0 indicates no correlation and 1 indicates they are fully correlated.

Feature clusters

The initial number of feature cluster is chosen as C and the first pivot feature is randomly selected. There is a corresponding feature group for each pivot that consists of all features that are closer to that pivot than to the other pivots. It is important to note that the effectiveness of feature grouping depends critically on the choice of the first pivots. Our selection criteria for pivot features are number c features with strong correlation. Every iteration continuously updates the pivot feature of each cluster. The algorithm stops when there is no update in the pivot feature in iteration. As a result, the closely correlated features are assigned to one of the feature cluster C_i, in a set of feature cluster C and the algorithm for feature clustering is given by Algorithm 1.

Algorithm 1: Feature clustering

1. Initialize the number of feature clusters, C
2. Choose a random feature as pivot and the remaining C-1 pivot features are selected by finding features with minimum feature correlation.
3. Assign the features to C clusters by placing strongly correlated features in a single cluster.
4. Repeat 2 and 3 until no pivot feature is updated

Feature weighting

The features in a feature cluster may have varied importance. Feature having large feature weight have more significance than the features with lesser feature weight. The feature weights are measured by determining average feature correlation among all the other features in the cluster.

If CDS represents a high-dimensional categorical dataset with n objects, y_i is the feature in feature cluster C_r which has totally m features. Then, the feature weight of feature y_i is measured as the average feature correlation between y_i and all the other features in the cluster C_r. Thus, the feature weight wt (y_i) of feature y_i ranges between 0 and 1 is given by Equation 6.

$$wt(y_i) = \frac{1}{p}\sum_{j=1}^{p} V(y_i, y_j) \quad (6)$$

Outlier scoring

The outlier score of an object x_i in feature cluster C_r is determined by the sum of the frequencies of x_i in all the features. So, the outlier score of x_i in CDS represented as Score(x_i) and is given as,

$$score(x_i) = \frac{1}{p}\sum_{j=1}^{p} (wt(y_j) * g(n(x_{i,j}))) \quad (7)$$

Where $x_{i,j}$ refers to the value that appears in the j^{th} feature of object x_i; $n(x_{i,j})$ refers to the frequency of x_i and g(x) is a constructed function which is equal to (x-1)log(x-1)-x log x[22].

The proposed algorithm constructs a total of C clusters for the dataset CDS and determines k outliers for every feature cluster. Hence, a total of C x k outliers is detected from all the feature groups. The outliers chosen from the C clusters are merged to a large set and is referred as candidate set of outliers. Therefore, the candidate set of outliers is the union of all the outlier sets from the C feature clusters.

The summary of the proposed Feature Grouping method is described in algorithm 2.

Algorithm 2: proposed feature grouping method

1. Determine feature correlation among the attributes
2. For every feature, determine the feature weight using Equation. 6
3. The outlier score of data object is determined with respect to each of the c clusters using Equation 7.
4. The outlier set of every cluster is built by determining k objects with highest outlier scores
5. The Candidate set of outliers is the union of the outlier sets of c feature clusters.

Experimental results

This section discusses the performance of the proposed outlier detection method evaluated on the UCI datasets and the comparison of the same with algorithms like LOF and Isolation Forest. The performance is measured by metrics like Precision, recall, F1-score and AUC.

Dataset description

The UCI datasets (Asuncion & Newman, 2007) are used for the experimental evaluation and is listed in Table 1. The results obtained are compared with LOF, Isolation Forest and WATCH in terms of precision, recall, F1-score and AUC.

Table 1. UCI Datasets.

Datasets	#Features	#Categorical features	#Data instances	Outlier%
Arrhythmia	279	73	452	4.58
Breast Cancer	9	5	286	5.6
Heartdisease	13	1	303	4.4
Lymphography	18	6	148	4.05
Mammography	5	2	961	46.3

The arrhythmia is a multiclass dataset with dimensionality 279, in which the smallest of the classes like 3, 4, 5, 7, 8, 9, 14 and 15 are merged to form the outliers and the other classes are merged to form the inliers. The breast cancer dataset includes raw data representing the level of cancer and their possibilities of occurrence. This dataset has two classes. The first class has 201 instances and the second class has 85 instances, with a total of 286 instances. These instances have 9 features of which some of them are categorical and some are nominal and a class attributes. The class attribute may be either no-recurrence-events or recurrence-events. The other features are age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad and irradiat. Heartdisease dataset is a multivariate dataset for detecting heart disease. In this, the attributes may be categorical, integer or real. Lymphography dataset is also a multi-class dataset with 4 classes, in which two classes containing very few data instances are considered as outliers and the other two classes are combined to form the inliers. In Mammography dataset, outliers are the minority class with calcification and the rest are considered as inliers.

Results and discussion

A heat map of the feature correlation between various features and the feature weights of all the categorical features is represented in a scale [0,1] with the darker color on the map indicating lower correlation between features and lighter shade indicating high correlation. Feature weights of different features are measured by quantifying the correlations among all the features in the group.

The performance of the outlier detection algorithm can be measured using various performance metrics like Precision, Recall, F1-score and AUC. Finally, the results are computed for LOF, Isolation Forest, WATCH and the proposed feature grouping algorithm.

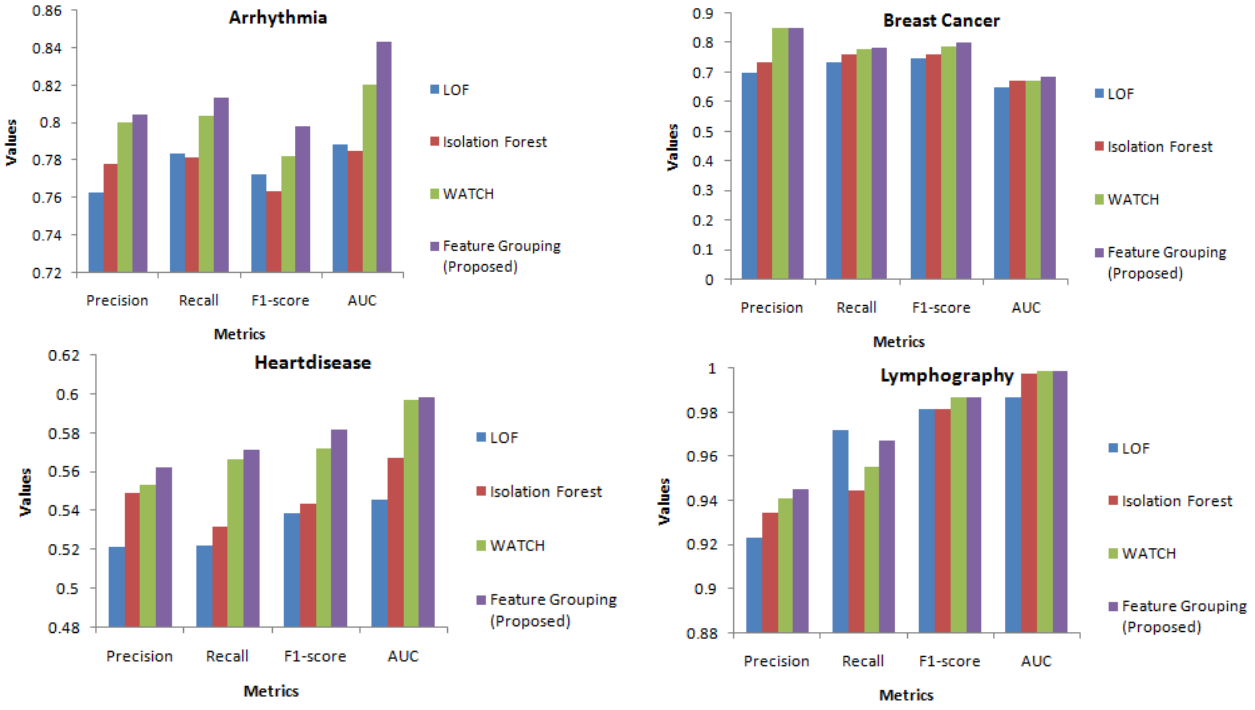
Precision is the ratio of true positives to all positives detected, whereas recall refers to the percentage of true positives to the total positives. F1-score represents the harmonic mean of precision and recall, considering both false positives and false negatives. The ROC (Receiver Operating Characteristics) curve is plotted with the True Positive Rate against the False Positive Rate. From the ROC curve the Area Under Curve (AUC) is calculated. Higher the value of the AUC, better is the performance of the algorithm.

Table 2 provides the comparison results of Feature grouping with traditional algorithms like LOF and Isolation Forest and the state-of-the-art algorithm, WATCH. From the results, it is observed that feature grouping algorithm provides a better performance when compared to the existing algorithms. The grouping of features based on the correlation between them and also considering the categorical attributes helps in improvising the performance metrics

Table 2. Performance Comparison of Feature Grouping with LOF and Isolation Forest for UCI datasets.

UCI Datasets	Metrics	LOF	Isolation Forest	WATCH	Feature Grouping (Proposed)
Arrhythmia	Precision	0.7623	0.7781	0.7998	0.8045
	Recall	0.7834	0.7812	0.8035	0.8131
	F1-score	0.7721	0.7634	0.7821	0.7982
	AUC	0.7885	0.7850	0.8202	0.8432
Breast Cancer	Precision	0.6975	0.7343	0.85	0.8523
	Recall	0.7332	0.7625	0.7769	0.7815
	F1-score	0.7492	0.7612	0.7892	0.8004
	AUC	0.6507	0.6708	0.6733	0.6832
Heartdisease	Precision	0.5212	0.5489	0.5528	0.5623
	Recall	0.5221	0.5312	0.5662	0.5711
	F1-score	0.5387	0.5432	0.5721	0.5814
	AUC	0.5458	0.5671	0.5970	0.5985
Lymphography	Precision	0.9234	0.9342	0.9412	0.9454
	Recall	0.9721	0.9446	0.9551	0.9674
	F1-score	0.9812	0.9815	0.9870	0.9865
	AUC	0.9870	0.9976	0.9989	0.9986
Mammography	Precision	0.8567	0.8675	0.8688	0.8767
	Recall	0.8898	0.8932	0.8977	0.8993
	F1-score	0.9145	0.9212	0.9393	0.9445
	AUC	0.9203	0.9402	0.9423	0.9529

Figure 2 explains the comparison of the results obtained for Feature grouping with the traditional algorithms with respect to Precision, recall, F1-score and AUC. This indicates that the proposed algorithm is able to detect the outliers effectively even after considering the categorical attributes.



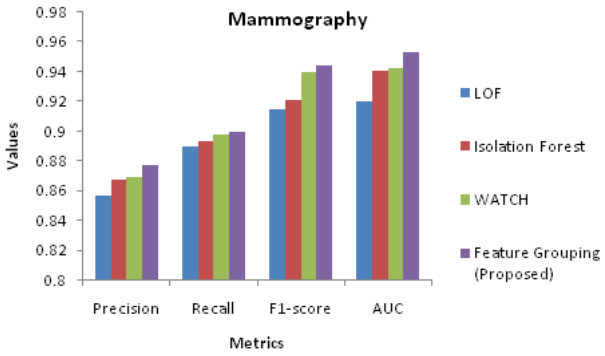


Figure 2. Comparison of Feature grouping with other algorithms for the UCI datasets

Conclusion

This paper presents the significance of a new algorithm that implements feature grouping for detecting outliers in categorical data. The novelty of the proposed method is that all the features of the objects in the dataset are checked for their inter-correlations and are grouped into feature groups based on the correlations and the closely correlated features are placed into a single group. The features are also assigned feature weights that project their importance and plays an important role in the outlier scoring. Finally, the outliers are detected and classified by assigning each data point as an outlier score. The traditional algorithms namely LOF and Isolation Forest and the state-of-the-art algorithm, WATCH are implemented to evaluate the performance of the feature grouping algorithm. Since LOF and Isolation Forest does not work with categorical data, Label Encoding is applied to transform the categorical data to integer values. The experimental evaluation was carried out using the UCI datasets. The outcome of experimental analysis shows that feature grouping algorithm outperforms the existing algorithms in terms of their performance metrics.

References

Asuncion, A., & Newman, D. (2007). *UCI machine learning repository*. Irvine, CA: Irvine University of California

Angiulli, F., Fassetti, F., Palopoli, L., & Serrao, C. (2022). A density estimation approach for detecting and explaining exceptional values in categorical data. Applied Intelligence, *52*(15), 17534-17556. DOI: https://doi.org/10.1007/s10489-022-03271-3

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (p. 93-104). Texas: ACM

Cao, F., Wu, X., Yu, L., & Liang, J. (2021). An outlier detection algorithm for categorical matrix-object data. *Applied Soft Computing*, *104*, 107182.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 1-58.

Cheng, L., Wang, Y., & Ma, X. (2019). A neural probabilistic outlier detection method for categorical data. *Neurocomputing*, *365*, 325-335.

Du, H., Ye, Q., Sun, Z., Liu, C., & Xu, W. (2020). FAST-ODT: A Lightweight Outlier Detection Scheme for Categorical Data Sets. *IEEE Transactions on Network Science and Engineering*, *8,* 13-24. DOI: https://doi.org/10.1109/TNSE.2020.3022869

Eiras-Franco, C., Martínez-Rego, D., Guijarro-Berdiñas, B., Alonso-Betanzos, A., & Bahamonde, A. (2019). Large scale anomaly detection in mixed numerical and categorical input spaces. *Information Sciences*, *487*, 115-127.

Femi, P. S., & Vaidyanathan, S. G. (2018). Comparative study of outlier detection approaches. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)* (p. 366-371). Coimbatore, IN: IEEE.

Femi, P. S., Vaidyanathan, S. G., & Kala, A. (2021). Integrating fuzzy constraint with feature correlation for local outlier mining. *Sādhanā, 46*(3), 1-11.

Femi, P. S., Vaidyanathan, S. G. (2022). An efficient ensemble framework for outlier detection using bio-inspired algorithm. *International Journal of Bio-Inspired Computation*, *19*(2), 67-76.

Garchery, M., & Granitzer, M. (2018). On the influence of categorical features in ranking anomalies using mixed data. *Procedia Computer Science*, *126*, 77-86.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London, UK: Chapman and Hall.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, *22*(2), 85-126.

Hu, X., Wang, Y., & Cheng, L. (2019). Multi-Hierarchy Attribute Relationship Mining Based Outlier Detection for Categorical Data. In *2019 International Joint Conference on Neural Networks (IJCNN)* (p. 1-8). Budapest, HU: IEEE.

Ienco, D., Pensa, R. G., & Meo, R. (2016). A semisupervised approach to the detection and characterization of outliers in categorical data. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(5), 1017-1029.

Knorr, E. M., & Ng, R. T. (1998). Algorithms for mining distance-based outliers in large datasets. In *VLDB. 98*, 392-403.

Lazarevic, A., & Kumar, V. (2005). Feature bagging for outlier detection. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (p. 157-166). Chicago IL: ACM

Li, J., Zhang, J., Pang, N., & Qin, X. (2018). Weighted outlier detection of high-dimensional categorical data using feature grouping. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *50*(11), 4295-4308.

Li, J., Zhang, J., Qin, X., & Xun, Y. (2019). Feature grouping-based parallel outlier mining of categorical data using spark. *Information Sciences*, *504*, 1-9.

Li, S., Shao, M., & Fu, Y. (2018). Multi-view low-rank analysis with applications to outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *12*(3), 1-22.

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. *In 2008 Eighth IEEE International Conference on Data Mining* (p. 413-422). Pisa, IT: IEEE.

Nowak-Brzezińska, A., & Łazarz, W. (2021). Qualitative data clustering to detect outliers. *Entropy*, *23*(7), 869.

Pai, H. T., Wu, F., & Hsueh, P. Y. (2014). A relative patterns discovery for enhancing outlier detection in categorical data. *Decision Support Systems*, *67*, 90-99.

Panjei, E., Gruenwald, L., Leal, E., Nguyen, C., & Silvia, S. (2022). A survey on outlier explanations. *The VLDB Journal*, *31*(5), 977-1008. DOI: https://doi.org/10.1007/s00778-021-00721-1

Shi, Y., & Zhang, L. (2011). COID: A cluster–outlier iterative detection approach to multi-dimensional data analysis. *Knowledge and Information Systems*, *28*, 709-733.

Tang, B., & He, H. (2017). A local density-based approach for outlier detection. *Neurocomputing*, *241*, 171-180.

Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in outlier detection techniques: A survey. *IEEE Access*, *7*, 107964-108000.

Wang, Y. F., Jiong, Y., Su, G. P., & Qian, Y. R. (2019). A new outlier detection method based on OPTICS. *Sustainable Cities and Society*, *45*, 197-212.

Zhang, J., Yu, X., Li, Y., Zhang, S., Xun, Y., & Qin, X. (2016). A relevant subspace based contextual outlier mining algorithm. *Knowledge-Based Systems*, *99*, 1-9.

Zhang, J., Yu, X., Xun, Y., Zhang, S., & Qin, X. (2017). Scalable mining of contextual outliers using relevant subspace. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *50*(3), 988-1002.