ICIC Express Letters
Volume **18**, Number **4**, April **2024**

 ISSN 1881-803X
pp. **333–342**

INCREMENTAL SEMI-SUPERVISED APPROACH TO DETECTING OUTLIERS IN CATEGORICAL DATA

Govind Pole* and Pradeepini Gera

Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Andhra Pradesh, Guntur 522302, India
pradeepini_cse@kluniversity.in
*Corresponding author: 14303019@kluniversity.in

Received April 2023; accepted July 2023

Abstract. *The majority of studies create training subsets for building base classifiers via random sampling. As a result, their diversity cannot be guaranteed, which could result in a decline in categorization performance as a whole. The aim of this study is to enhance clustering outcome for categorical data by an ensemble of clustering and classification techniques in a semi-supervised environment. In order to achieve good results with categorical data, ad hoc strategies are used. The cluster-based incremental ensemble member selection method is used to choose the ensemble members. Each participating cluster is assessed for accuracy of clustering and compactness. To create several trainings subsets, it is then necessary to pairwise combine clusters from various classes. Each subgroup of the training set is given a unique base classifier. The outcomes of these base classifiers are combined by weighted voting for a sample whose class label needs to be predicted. A collection of distance-based techniques were used to discriminate between typical and anomalous occurrences after developing a model that captures the traits of normal instances. The experimental study uses seven benchmark data sets to assess performance. To evaluate the performance of an outlier detection, the Area under the Curve (AUC) scores are calculated. The performance shift from 5 to 20% is observed for different data when compared to the existing work. Thus, comparing the suggested procedure with traditional ensemble of clustering and classification algorithms, the results show that the Incremental Semi-Supervised Ensemble of Clustering and Classification (ISECC) can create ensembles with greater variation and improved accuracy.*

Keywords: Clustering ensemble, Classification ensemble, Semi-supervised approach, Outliers

1. **Introduction.** Any variable that accepts names, attributes, or labels is referred to as a categorical variable. For instance, a group of equities may be classified as either a growth stock or a value stock depending on how the stocks are invested. Investment style serves as a categorical indicator of whether a company falls under the growth or value category. Another option is to use a categorical variable to determine if a company has survived or gone bankrupt for all companies over a certain period, like the last 20 years. Financial econometrics frequently deals with issues where the dependent variable, one or more explanatory factors, or both are categorical variables. To handle categorical variables, specific regression models or methods are available. When one or more of the explanatory variables are categorical, regression analysis with dummy variables is a technique utilized. The three models (together referred to as probability models) that can be employed when the dependent variable is a categorical variable are the linear probability model, the logit regression model, and the probit regression model [1]. In many fields, such as social networking sites, forums for comments and reviews, genetic analysis of various

DOI: 10.24507/icicel.18.04.333

animals, biology, chemistry, bioinformatics, genomics, and medicine, among others, categorical data is widely used. Age, sex, caste, religion, DNA sequence data, and molecular biology data are a few examples of the country's population information. Several intrinsic problems with this type of data include lack of natural ordering, high dimensionality, sparse distribution, and lack of arithmetic characteristics. As there is no generic method for analyzing categorical data, it requires a unique consideration for effective analysis. After replacing the various categories, symbols, and words with numbers, some numerical analysis techniques are used. Examples of this qualitative to quantitative conversion include contrast coding, effect coding, non-sense coding, and dummy coding [2]. Clustering has been attempted to be semi-supervised incorporated into the classification model in a variety of ways. Popular techniques include try-training, ASSEMBLE, Semi Boost, and more. Their main objective was to learn from a lot of unlabeled data and a little of labelled data. Less than we did, they just took account of one basic classifier and one base clustering method. The training set was divided into various clusters, and these clusters of pairwise classes were then concatenated to produce multiple training examples. Each classifier is trained on a specific training sample, and the results are combined using a weighted voting method [3].

Unusually, behaviors in industrial systems may be early indicators of crucial occurrences that could seriously harm infrastructure and security. So, it is crucial to precisely and promptly identify anomalous actions. Yet, the difficulty of solving the anomaly detection problem in practice is mostly owing to the rarity and high expense of obtaining labels for abnormalities [4]. Data patterns that are distinct from the existing data are revealed through outlier identification. The outlier identification method for numerical datasets based on k-nearest neighbor network has garnered a lot of attention in recent years thanks to its good robustness and interpretability. The datasets created in many practical situations, or datasets with mixed-valued attributes, however, tend to include both numerical and categorical variables. With unlabeled datasets, choosing the appropriate value of k is another problem that needs to be addressed [5]. The unsupervised outlier identification research, which uses an unlabeled data set with abnormality assumptions, has received a lot of attention. We suggest transferring the knowledge from the labelled source data to the target data set to help with the unsupervised outlier detection because the target data set has a wealth of associated labelled data that is available as auxiliary information. The source data and target data are combined for joint clustering and outlier detection using the source data cluster structure as a constraint in order to fully utilize the source knowledge. To do this, the destination data's partitions are regularized using the categorical utility function so that they are consistent with the labels of the source data. Using an augmented matrix and a K-means technique, the issue is completely handled with a precise mathematical description and theoretical convergence guarantees. We applied incremental semi-supervised algorithms to nine real-world data sets for thorough testing and comparison. The results demonstrate the effectiveness of the offered solutions and demonstrate a considerable improvement in metrics for outlier detection and cluster validity [6]. Anomaly detection, also known as outlier identification, seeks to locate the minority data points that stand out from the majority in a variety of real-world contexts, such as credit card fraud, network infiltration, precision marketing, and other things. Considerable advances have been made in this area, particularly in the area of unsupervised outlier detection. Many approaches are proposed from various mathematical angles, including density-based local outliers [7], local distance-based outlier detection [8], angle-based outlier detection [9], ensemble-based isolation forest [10], anomaly detection using principal component analysis [11], and so on.

2. Related Work.

2.1. **Ensemble classification methodologies.** Using partially labelled data, Zhong et al. [12] suggested a semi-supervised multiple-choice learning strategy to jointly train a network ensemble. It emphasizes enhancing the assignment of labelled data among the individual networks and utilizing unlabeled data to obtain domain-specific knowledge. While minimizing the conditional entropy regarding the posterior probability distribution, it adopts a negative 1-norm regularization. Liu et al. [13] offered a selective ensemble learning method for BRB Classification Systems (BRBCS) based on multi-objective Pareto Archived Evolutionary Strategy (PAES) optimization. Using the improved bagging algorithm, it learns the base classifier. To raise the level of difference in the basic classifier's integration, the training set is produced by repeatedly sampling data. The number of base classifiers involved in the base classifier's integration and generalization error are utilized as the objective functions for multiobjective optimization, and the trained base classifier is binary coded during the base classifier selection stage. To strike a compromise between accuracy and diversity, Bian et al. [14] developed ensemble pruning based on objection maximization. It formalizes the ensemble pruning problem as an information entropy-based objection maximization problem. It suggests an ensemble pruning technique with both a centrally managed version and a distributed version, the latter of which is meant to accelerate the former. Yang et al. [15] presented multi-instance's ensemble learning using discriminative bags. For the multi-instance Ensemble Learning with Discriminative Bag (ELDB) algorithm, it provides two unique approaches. Two sections claim that the bag selection method creates a discriminative Bag Set (dBagSet). In order to construct the fundamental dBagSet while taking consideration of the space and label distribution of the data, the bag selection method is first optimized using discriminative analysis. A dBagSet with greater distinguishability can be produced via self-reinforcement using the state and action transfer technique, as well. The ensemble technique trains a number of classifiers using these dBagSets to produce the final weighted model.

2.2. **Outlier detection and clustering.** In the field of data mining, cluster analysis and outlier identification are two areas that are constantly gaining attention. Clustering with Outlier Removal (COR) takes account of both the cluster analysis and the problem of outlier detection. Here, by creating fundamental partitions, the original space is converted into a binary space. An auxiliary binary matrix is provided for a clean and effective solution, allowing COR to fully and effectively tackle the difficult problem using a unified K-means with theoretical backing [16].

The rapid influx of data in data streams necessitates quick computation in the least amount of time and memory. To reduce the computing cost of effectively finding distance-based outliers, the author here offers a new distance-based outlier detection methodology. The two methodologies make up the proposed Micro-Cluster with Minimal Probing (MC-MP) technique. First, it uses micro-clusters to reduce the need for range queries. The concept of separating strong and trivial inliers is then proposed to deal with the items outside of the micro clusters [17]. On the foundation of the idea that effective data compression will encode outliers with distinctive symbols, rate-distortion theory-based outlier identification is developed. It suggests two effective algorithms for cluster purging, one of which has no parameters and the other of which has a parameter that regulates representivity estimates, allowing it to be adjusted in supervised settings [18]. Only potential outliers are permitted to have membership values lower than "1" in a newly proposed method for allocating membership values. In order to accomplish this, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering is first used to identify a collection of potential outliers, and these potential outliers were then given membership values based on a few heuristics. The membership value of "1" was given to all other

remaining samples [19]. Here, in order to deal with data with imperfect labels and incorporate a small number of atypical cases into learning, the author introduces a novel outlier identification strategy. In order to handle data with imperfect labels, we propose probability values for each input data set, which indicate, respectively, the degree to which each example belongs to the normal and abnormal classes. The suggested method operates in two steps. By computing each example's probability values based on its local behaviour, it first creates a fictitious training dataset. The kernel k-means clustering method and the kernel LOF-based method are used to compute the likelihood values. It incorporates the derived probability values and restricted aberrant examples into a Support Vector Data Description (SVDD) based learning architecture to provide a classifier for global outlier detection that is more accurate [20]. Only a few known methods can handle categorical data, despite the fact that there are numerous outlier detection algorithms for applications to numerical data. High temporal complexity and low detection precision are two important issues that plague categorical data method design. Here, the author offers two brand-new categorical data set outlier identification algorithms. It begins by outlining a simple entropy-based strategy known as the outlier detection tree. The data set is split into two categories by Outlier Detection Tree (ODT) using a classification tree: a normal group and an aberrant group. Each data object is then labelled as an outlier or a regular one using the if-then rule in the tree. Moreover, we offer FAST-ODT, a powerful outlier identification method with minimal time complexity and good detection accuracy [21]. This article's author recommends possibilistic exponential fuzzy clustering, which lessens and completely eliminates the influence of outliers throughout the clustering process [22].

From the previous work, it is observed that cluster analysis and outlier identification have a close relationship with one another and the majority of current studies treat these two activities independently rather than recognizing the linked relationship between them in real life. Only a few known methods can handle outliers in categorical data. The majority of the discussed techniques do not practice the advantage of clustering and classification association. The synthesis of clustering and classification in a semi-supervised environment is gaining attention of researchers. To reduce the computing cost, the cluster purging emerges as an essential step towards the optimality of analysis.

3. **Proposed System.** The following four well-known heuristics form the foundation of the suggested system.

1) Similarity between the class distributions of a group and its individual members: If an object is a member of a group, then both the object and the group's class distributions should be comparable.

2) Similarity between two items within a group: The likelihood that two objects are members of the same class increases if they are allocated to the same group. It is known as the "co-occurrence principle".

3) Similarity between the object's final class distribution and its average class distribution: An object's final class distribution ought to be more similar to the average class distribution that the base classifiers were able to determine. This idea is known as the "consensus principle".

4) Similarity between the group's final class distribution and its average class distribution of constituent objects: A group's final class distribution should be more similar to the average class distribution of its component objects.

The proposed incremental semi-supervised ensemble of clustering and classification is depicted in Figure 1. The large categorical dataset is divided into multiple subsets and distributed to different machine (slave) through coordinators. All clusters created by each participating node are evaluated for compactness and quality of clustering. The best value of K (number of clusters) is selected from all results by the coordinator machine. The best value of K is used for further processing. The semi-supervised clustering is performed with

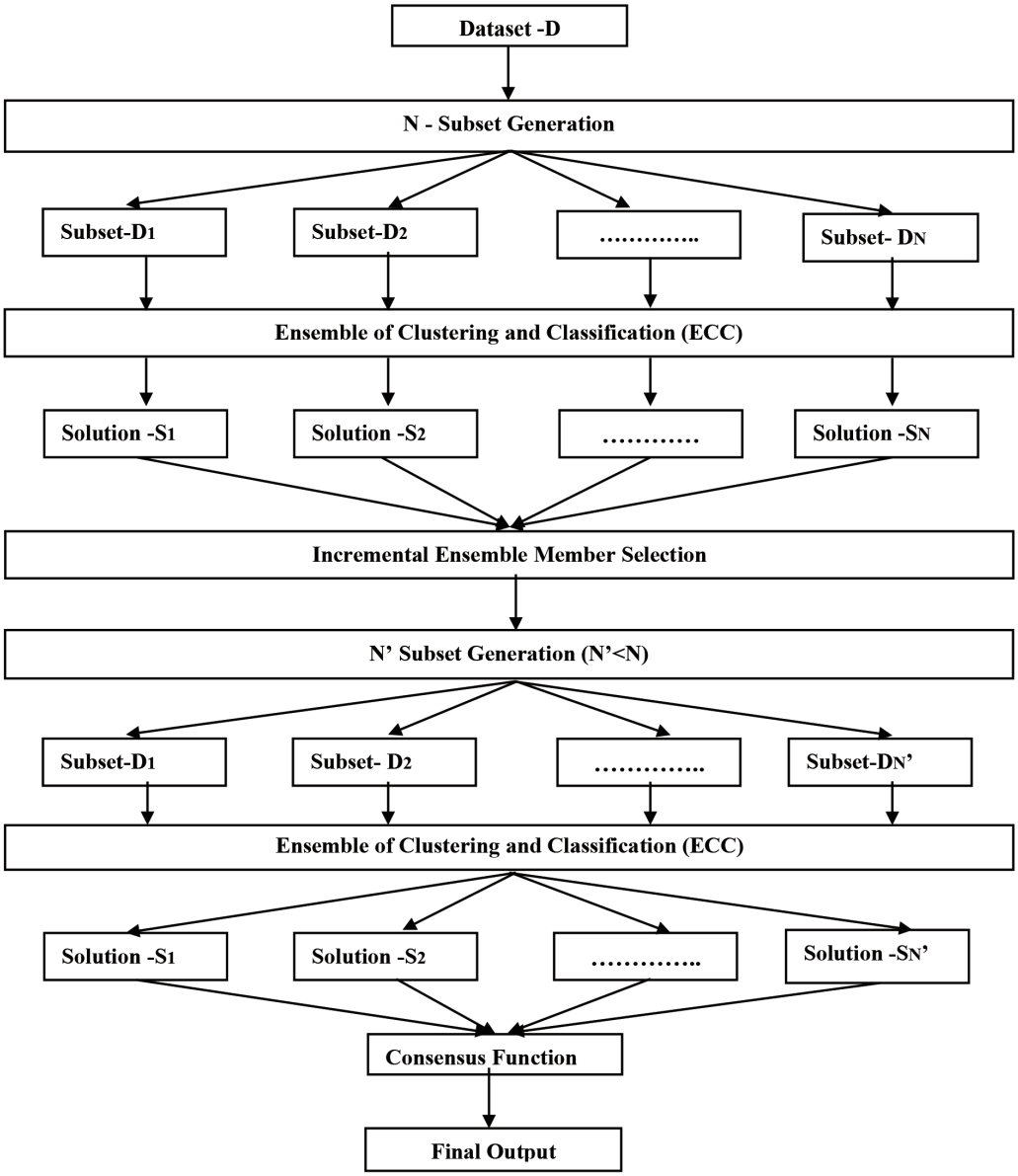

FIGURE 1. Incremental semi-supervised ensemble of clustering and classification

predetermined value of K for cluster optimization using Ensemble Clustering and Classification (ECC). For the production of new ensembles, Cluster Based Incremental Ensemble Member Selection (CBIEMS) is required. Moreover, the newly produced ensemble is used to perform Ensemble Clustering and Classification (ECC). The newly discovered clusters are combined to create the optimal cluster. This is achieved with the help of the proposed Incremental Semi-Supervised Ensemble of Clustering and Classification (ISECC).

Algorithm 1: Incremental Semi-Supervised Ensemble of Clustering and Classification (ISECC)

Input: High dimensional dataset D

Process:

1. Creation of the initial ensemble.
2. Create "N" random subspaces, such as $A_1, A_2, \ldots, A_N$.
3. Using clustering and classification, create the semi-supervised models $M_1, M_2, \ldots, M_N$.
4. Invoke Algorithm 2's incremental ensemble member selection procedure.
5. A new generation of ensembles.

6. Create $A_1, A_2, \ldots, A_{N'}$ randomly chosen subspaces, where $(N' < N)$.
7. Use Clust+Class to create semi-supervised clustering models $M_1, M_2, \ldots, M_{N'}$.
8. Compile the Clust+Class solutions $S_1, S_2, \ldots, S_{N'}$ produced by the semi-supervised models to obtain the consensus matrix O.
9. Consensus function employing the normalized cut method for the final findings;

Output: The labels of the samples in D.

Algorithm 2: Cluster Based Incremental Ensemble Member Selection (CB-IEMS)

Input:

Random Subsets $A = \{A_1, A_2, \ldots, A_N\}$.
Semi-supervised clustering models $M = \{M_1, M_2, \ldots, M_N\}$.
Ensemble members $E' = \{(A_1, M_1), (A_2, M_2), \ldots, (A_N, M_N)\}$.
The empty ensemble $E = \{\ \}$.

Process:

1. For n in $1, 2, \ldots, N$.
2. Calculate the objective function F_N for each clustering solution S_N generated by ensemble
3. Set $t = 1$
4. Sort ensemble members in ascending order according to the corresponding F_N, and pick up the first ensemble member (A_t, M_t);
5. Add to new ensemble: $E = \{(A_t, M_t)\}$, $E' = E - \{(A_t, M_t)\}$
6. Repeat
7. $t = t + 1$;
8. For each (A_n, M_n) in E'
9. Calculate the local objective function L_N.
10. Sort the ensemble members in E' in ascending order according to the corresponding local objective function L_N.
11. Set $n = 0$
12. Repeat
13. Set $n = n + 1$;
14. New ensemble $E'' = E + \{(A_n, M_n)\}$, where $\{(A_n, M_n) \in E'\}$
15. Calculate the global objective function $F_{N''}$ and F_N for the clustering solutions $S_{N''}$ and S_N respectively
16. Until $F_{N''} <= F_N$;
17. Add to new ensemble: $E = E + \{(A_n, M_n)\}$, $E' = E' + \{(A_n, M_n)\}$.
18. Until $t >= N'$ or $E' = \Phi$

Output: The new ensemble E.

4. **Results and Discussion.** In this section, brief details of experimental setup are presented along with a collection of base classifiers and base clustering techniques whose outputs are merged, and the experimentation-related algorithms. Seven datasets total, all acquired from the common UCI machine learning repository, are used. In Table 1, a summary of these datasets is presented. Our base classifiers are trained using this division. Base clustering techniques, however, are applied to the entire dataset. Our strategy uses the results of the base classifier and base clustering techniques on just the test dataset. We must take account of both the detection rate (the number of instances of the abnormal class found by the algorithm) and the detection error when assessing the performance of an outlier identification system (the amount of instances of the normal class that the algorithm misjudges as outliers). Usually, the Area under the Curve (AUC), which takes account of both measurements, is employed to assess the outcomes [30]. In this study, we follow the methodology suggested in [31], where the AUC score is calculated using a

Table 1. AUC score for different ensemble approaches

Datasets	Local Outlier Factor (LOF)	unconstrained Least-Square Importance Fitting (uLSIF)	One Class SVM (OSVM)	Feature ensemble model (FRaC)	Proposed ISECC
Mushroom	0.448	0.3701	0.5961	0.551	**0.756**
Breast Cancer	0.509	0.362	0.526	0.526	**0.705**
Dermatology	0.785	0.358	0.889	**0.895**	0.860
Hepatitis	0.648	0.361	0.814	0.856	**0.892**
Nursery	0.585	0.359	0.566	0.581	**0.787**
Adult	0.448	0.3701	**0.596**	0.551	0.574
Credit-Approval	0.520	0.357	0.731	0.467	**0.799**

closed-form formula,

$$AUC = [S_0 - n_0(n_0 + 1)/2]/n_0 n_1$$

where, n_0 represents the quantity of test examples that belong to the normal class and n_1 represents the quantity of abnormal test instances and $S_0 = \sum_{i=1}^{n_0} r_i \in r_i$, where r_i is the rank that the ith normal instance in the test set is assigned by the normal class's class model. That is, the OS score assigned to each normal instance in the test set in our example.

We compared our proposed method ISECC with four existing works, LOF [7], OSVM [27], uLSIF [28], and FRaC [29]. We must pair the LOF with a distance function that can handle this kind of data in order to enable the LOF to work with categorical attributes. We chose to link LOF with the occurrence frequency distance function because it was claimed that this metric produced the best performance outcomes. Mismatches with unusual values receive a high distance value under this metric. Ten equi-depth bins per data set were used to discretize the numerical aspects of each data set. We translated each category attribute, gave each categorical value a Boolean attribute, and finished the data pre-processing required for uLSIF and OSVM (standard pre-processing for SVM). We apply the same pre-processing to uLSIF.

The experimental findings are detailed in this section. Numbers of experiments are performed on actual categorical data sets, followed by a comparison with the prior work, to evaluate the effectiveness of the proposed methodology. In Table 1 experimental findings are reported. To evaluate the performance of an outlier detection, the Area under the Curve (AUC) score is calculated. The AUC considers both the detection rate and the detection error. Thus, the measures for the amount of instances of the abnormal class found by the algorithm and the amount of instances of the normal class that the algorithm misclassifies as outliers are considered simultaneously. The bold face values in Table 1 indicate winning score for the dataset. In most of the cases, ISECC triumphs. Figure 2 compares the existing works, the AUC score of ISECC outperforms for the datasets mushroom, breast cancer, hepatitis, nursery, credit-approval. The performance shift is from 5 to 20% as compared to the existing work. Here for two datasets dermatology and adult, the AUC score of the proposed approach is less than the existing work, but it is still comparable to the winner's AUC score.

Seven datasets total (Table 2), all acquired from the common UCI machine learning library, are used in this study [23].

5. **Conclusion and Future Scope.** Categorical data management and processing is a common issue in data mining. This type of data typically requires ad hoc procedures in

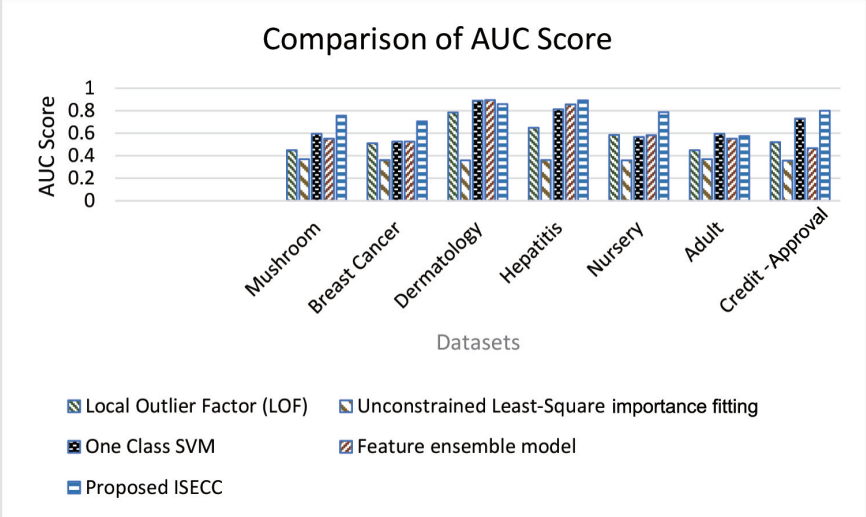

FIGURE 2. Comparison of AUC score

TABLE 2. Dataset details

Datasets	Instances	Attribute	Attribute type	Classes	Class distribution
Mushroom	8124	22	Categorical	2	4208, 3196
Breast cancer	286	9	Categorical	2	201, 85
Dermatology	366	33	Categorical, Integer	6	112, 61, 72, 49, 52, 20
Hepatitis	155	19	Categorical, Integer, real	2	32, 123
Nursery	12960	8	Categorical	5	4320, 2, 328, 4266, 4044
Adult	48842	14	Categorical, Integer	2	NA
Credit-Approval	690	15	Mixed	2	307, 383

order to produce appropriate findings. Hence, a novel strategy based on ISECC is presented in this study. Categorical data can be modelled using the proposed framework and a distance-based approach. In comparison to other cutting-edge semi-supervised algorithms for anomaly identification, we achieved quite good results.

Incremental semi-supervised clustering ensemble improves the accuracy and efficiency of the clustering mechanism. The hierarchical approach insures the systematic delegation and coordination of work. The specifically developed methods outperform the general-purpose methods for the administration of categorical data. It detects outliers in the categorical data where no general measure is available to data analysis. For future work, we consider the analysis of real-time data such as stock share price trends, news analysis, social media data analysis in more real time by reducing the time interval for data analysis using the parallel and distributed computing.

REFERENCES

[1] F. Cao, J. Liang, L. Bai, X. Zhao and C. Dang, A framework for clustering categorical time-evolving data, *IEEE Transactions on Fuzzy Systems*, vol.18, no.5, pp.872-882, DOI: 10.1109/tfuzz.2010.2050891, 2010.

[2] J. Ji, W. Pang, Z. Li, F. He, G. Feng and X. Zhao, Clustering mixed numeric and categorical data with cuckoo search, *IEEE Access*, vol.8, pp.30988-31003, DOI: 10.1109/access.2020.2973216, 2020.

[3] T. Chakraborty, EC3: Combining clustering and classification for ensemble learning, *Journal of LaTeX Class Files*, vol.13, no.9, 2014.
[4] Q. Xie, P. Zhang, B. Yu and J. Choi, Semisupervised training of deep generative models for high-dimensional anomaly detection, *Transactions on Neural Networks and Learning Systems*, vol.33, no.6, pp.2444-2453, DOI: 10.1109/tnnls.2021.3095150, 2022.
[5] Y. Wang, X. Cao and Y. Li, Unsupervised outlier detection for mixed-valued dataset based on the adaptive k-nearest neighbor global network, *IEEE Access*, vol.10, pp.32093-32103, DOI: 10.1109/access.2022.3161481, 2022.
[6] W. Yu, Z. Ding, C. Hu and H. Liu, Knowledge reused outlier detection, *IEEE Access*, vol.7, pp.43763-43772, DOI: 10.1109/access.2019.2906644, 2019.
[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, LOF: Identifying density-based local outliers, *Proc. of SIGMOD*, vol.29, no.2, pp.93-104, 2000.
[8] K. Zhang, M. Hutter and H. Jin, A new local distance-based outlier detection approach for scattered real-world data, in *Advances in Knowledge Discovery and Data Mining. PAKDD 2009. Lecture Notes in Computer Science*, T. Theeramunkong, B. Kijsirikul, N. Cercone and T. B. Ho (eds.), Berlin, Heidelberg, Springer, 2009.
[9] N. Pham and R. Pagh, A near-linear time approximation algorithm for angle-based outlier detection in high-dimensional data, *Proc. of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*, pp.877-885, 2012.
[10] F. T. Liu, K. M. Ting and Z.-H. Zhou, Isolation forest, *Proc. of the 8th IEEE International Conference on Data Mining (ICDM)*, Pisa, Italy, pp.413-422, DOI: 10.1109/ICDM.2008.17, 2008.
[11] Y.-J. Lee, Y.-R. Yeh and Y.-C. F. Wang, Anomaly detection via online oversampling principal component analysis, *IEEE Transactions on Knowledge and Data Engineering*, vol.25, no.7, pp.1460-1470, 2013.
[12] J. Zhong et al., Semisupervised multiple choice learning for ensemble classification, *IEEE Transactions on Cybernetics*, vol.52, no.5, pp.3658-3668, DOI: 10.1109/TCYB.2020.3016048, 2022.
[13] W. Liu, W. Wu, Y. Wang, Y. Fu and Y. Lin, Selective ensemble learning method for belief-rule-base classification system based on PAES, *Big Data Mining and Analytics*, vol.2, no.4, pp.306-318, DOI: 10.26599/Bdma.2019.9020008, 2019.
[14] Y. Bian, Y. Wang, Y. Yao and H. Chen, Ensemble pruning based on objection maximization with a general distributed framework, *IEEE Transactions on Neural Networks and Learning Systems*, vol.31, no.9, pp.3766-3774, DOI: 10.1109/tnnls.2019.2945116, 2020.
[15] M. Yang, Y.-X. Zhang, X. Wang and F. Min, Multi-instance ensemble learning with discriminative bags, *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol.52, no.9, pp.5456-5467, DOI: 10.1109/tsmc.2021.3125040, 2022.
[16] H. Liu, J. Li, Y. Wu and Y. Fu, Clustering with outlier removal, *IEEE Transactions on Knowledge and Data Engineering*, vol.33, no.6, pp.2369-2379, DOI: 10.1109/tkde.2019.2954317, 2021.
[17] M. J. Bah, H. Wang, M. Hammad, F. Zeshan and H. Aljuaid, An effective minimal probing approach with micro-cluster for distance-based outlier detection in data streams, *IEEE Access*, vol.7, pp.154922-154934, DOI: 10.1109/access.2019.2946966, 2019.
[18] M. B. Toller, B. C. Geiger and R. Kern, Cluster purging: Efficient outlier detection based on rate-distortion theory, *IEEE Transactions on Knowledge and Data Engineering*, vol.35, no.2, pp.1270-1282, DOI: 10.1109/tkde.2021.3103571, 2023.
[19] R. K. Sevakula and N. K. Verma, Clustering based outlier detection in fuzzy SVM, *2014 IEEE International Conference on Fuzzy Systems (Fuzz-IEEE)*, Beijing, China, pp.1172-1177, DOI: 10.1109/fuzz-ieee.2014.6891600, 2014.
[20] B. Liu, Y. Xiao, P. S. Yu, Z. Hao and L. Cao, An efficient approach for outlier detection with imperfect data labels, *IEEE Transactions on Knowledge and Data Engineering*, vol.26, no.7, pp.1602-1616, DOI: 10.1109/tkde.2013.108, 2014.
[21] H. Du, Q. Ye, Z. Sun, C. Liu and W. Xu, FAST-ODT: A lightweight outlier detection scheme for categorical data sets, *IEEE Transactions on Network Science and Engineering*, vol.8, no.1, pp.13-24, DOI: 10.1109/tnse.2020.3022869, 2021.
[22] K. Treerattanapitak and C. Jaruskulchai, Outlier detection with possibilistic exponential fuzzy clustering, *2011 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Shanghai, China, pp.453-457, DOI: 10.1109/fskd.2011.6019597, 2011.
[23] *Dataset Link*, http://archive.ics.uci.edu/datasets, 2023.
[24] M. Anila and G. Pradeepini, Study of prediction algorithms for selecting appropriate classifier in machine learning, *Journal of Advanced Research in Dynamical and Control Systems*, vol.9, no.18, pp.257-268, 2017.

[25] H. S. A. Bommadevara, Y. Sowmya and G. Pradeepini, Heart disease prediction using machine learning algorithms, *International Journal of Innovative Technology and Exploring Enginehering*, vol.8, no.5, pp.270-272, 2019.
[26] L. Bai et al., A novel attribute weighting algorithm for clustering high-dimensional categorical data, *Pattern Recognition*, vol.44, no.12, 2011.
[27] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola and R. C. Williamson, Estimating the support of a high-dimensional distribution, *Neural Computation*, vol.13, no.7, pp.1443-1471, 2001.
[28] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama and T. Kanamori, Statistical outlier detection using direct density ratio estimation, *Knowledge and Information Systems*, vol.26, no.2, pp.309-336, 2011.
[29] K. Noto, C. Brodley and D. Slonim, FRaC: A feature-modeling approach for semi-supervised and unsupervised anomaly detection, *Data Mining and Knowledge Discovery*, vol.25, no.1, pp.109-133, 2012.
[30] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, vol.30, no.7, pp.1145-1159, 1997.
[31] D. J. Hand and R. J. Till, A simple generalisation of the area under the ROC curve for multiple class classification problems, *Machine Learning*, vol.45, no.2, pp.171-186, 2001.