Received: 00 Month 0000 | Revised: 00 Month 0000 | Accepted: 00 Month 0000

DOI: xxx/xxxx

ARTICLE TYPE

An Enhanced Hidden Semi-Markov model for Outlier Detection in Multivariate Datasets

G. Manoharan[1] | K. Sivakumar[2]

[1]Research Scholar, Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu, India

[2]Professor, Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai, Tamilnadu India

Correspondence

*Corresponding author name. Email: siva111k@gmail.com Email: vijimanoharan77@gmail.com

ORCHID:0000-0003-3919-0967

Summary

Outlier detection in data mining is an important arena where detection models are developed to discover the objects that do not confirm the expected behavior. The generation of huge data in real time applications makes the outlier detection process into more crucial and challenging. Traditional detection techniques based on mean and covariance are not suitable to handle large amount of data and the results are affected by outliers. So it is essential to develop an efficient outlier detection model to detect outliers in the large dataset. The objective of this research work is to develop an efficient outlier detection model for multivariate data employing the enhanced Hidden Semi-Markov Model (HSMM). It is an extension of conventional Hidden Markov Model (HMM) where the proposed model allows arbitrary time distribution in its states to detect outliers. Experimental results demonstrate the better performance of proposed model in terms of detection accuracy, detection rate. Compared to conventional Hidden Markov Model based outlier detection the detection accuracy of proposed model is obtained as 98.62% which is significantly better for large multivariate datasets.

KEYWORDS:

Outlier detection, Multivariate data, Hidden Markov Model

1 | INTRODUCTION

Identifying the outliers is an essential process in data mining while handling large amount of multivariate data. Discovery of data pattern that deviates from the rest is termed as an outlier and it is essential to detect such outliers in the data as it might introduce serious issues in the final results as model misspecification, errors in parameter estimation, and incorrect results. Various outlier detection modelssuch as Distance-based[1], Density-based[2], Statistical based[3], Clustering-based[4], Graph-based[5], Ensemble-based[6] methods are introduced by researchersto identify outliers in the large dataset. But still, there is a considerable discussion being in progress to define the outliers. The rule of thumb to flag the data point as an outlier is used in many research works[7] and detects the outliers if the standard deviation of the data point has much deviation from the mean. However, it is not suitable for all the circumstances and it might be wrong for unstructured large dynamic data. So it is vital to frame some crucial questions on handling the outliers. Since the data with the outliers will introduce a negative impact on the results it is essential to bring an efficient outlier detection model. The major objective of this research work is summarized as follows.

- To obtain an efficient outlier detection model that is suitable for both univariate and multivariate data.
- To improve the detection accuracy and detection rate of the outlier detection model.

The research work is further organized as follows: Section 2 provides a brief analysis of existing relevant literature works, Section presents the proposed outlier detection model, Section 4 presents the experimental observations, and the conclusion is presented in section 5.

2 | RELATED WORKS

Various outlier detection models are introduced in recent times and this section provides an overview of techniques as a short literature analysis. The probabilistic outlier detection model[8] detects the outliers and inliers using a log bilinear neural model by learning the categorical distributions in the dataset. Based on the learning loss the inliers and outliers are identified. However, the probabilistic model needs more parameters to detect the outliers, also it requires a hypothesis to verify the results which incur computational complexity and uncertainty in decision. Efficient rate pattern based outlier detection reported in[9] detects the outliers based on mining the rare patterns from the incremental data. It is a modified version of the frequent pattern mining algorithm. Though the pattern-based outlier detection simple it faces difficulties while handling complex data and lags while handling diverse attributes.

Virtual outlier score obtained for the outliers by constructing a similarity graph[10] virtually that uses local information to connect the points in the graph. Further tailored Markov random walk process is used to improve the connectivity of the graph and the outliers are identified. However, establishing connectivity to all the data points virtually for large datasets increases the system complexity. Graph based method can produce false positive results in the outlier detection process as it ignores the local information. Considering the local information around each object an outlier detection model reported in[11] constructs a local information graph and calculates the outliers. A feature map is used in[12] to detect the outliers by calculating each data point in a cluster. Relative K-distance neighborhood is used to refine the outlier detection process with improved accuracy and minimum computation time.

The weighted outlier mining method reported in[13] detects the outliers by grouping the features based on correlation and assigns scores for objects in the feature group. The ranking process produces categorical data and separates the outliers as specified by the users. Another feature grouping and outlier mining model reported in[14] is a parallel outlier mining technique that detects the outliers in a high dimensional dataset. The technique is implemented in the spark platform and obtains better performance by handling the complex datasets. However, clustering based outlier detection models provide better results than feature grouping. Markov boundary-based outlier detection[15] predicts the outliers considering each attribute in the dataset. The framed boundary considers the subspace and detects the different behavior data points as outliers. Sparse coding based outlier detection model[16] detects the outliers without any application specific or domain specific information. However, it is suitable for low dimensional data with lower complexity.

Cross-correlation analysis[17] for outlier detection handles high dimensional dataset with assembled and isolated outliers. By converting the high dimensional data into one-dimensional cross-correlation function the isolated outliers are identified. The abnormal samples are identified at different levels by ranking the samples at different levels. Reduced computation time is the major merit of this method, however, the detection accuracy can be improved further. The geometric reasoning based detection model reported in[18] detects outliers and inliers based on logical argument and spatial reasoning. This distance-based approach computes the corresponding distances and forms a triangle. The data points that are outside the triangle are considered as outliers in the detection process. However, it requires more parameters to compute the distance and triangle which is the major setback of the research work.

The hierarchical partitioning-based outlier detection model reported in[19] generates a tree structure for the input dataset and measures the dissimilarity between the instances to detect the outliers. Outlier score is used to compute the degree of an outlier for each instance and attains better performance than distance-based detection methods. Fuzzy based outlier detection model reported in[20,21] detects the outliers based on the similarity between the objects. Fuzzy membership functions and sparse threshold concept is used to evaluate the detection model performance by detecting the true samples and outliers. Improved accuracy is the merits of the detection model however it requires more logical conditions to validate the results.

From the research analysis, it can be observed that traditional methods experience issues while handling multidimensional data. The necessity of extra parameters to detect the outliers, suitable for low dimensional data, increased computation complexity are the major setbacks of the traditional models. whereas supervised approaches need large training process which degrades the results. Considering this as research motivation, this research work proposed an efficient outlier detection model using the enhanced Hidden Semi Markov model to detect outliers in the univariate and multivariate dataset.

3 | PROPOSED WORK

The proposed outlier detection model is presented in this section. An enhanced Hidden Semi Markov Model (HSMM) is used for detecting the outliers in the dataset.The conventional Hidden Markov Model (HMM) is useful to model sequence symbols. Unobservable conditions of the system are represented as the states of HMM. It produces certain probability as outputs and next states using characteristic parameters. However, HMM experiences probability density state is exponential which might affect the performance. To overcome this limitation, the Hidden semi Markov Model (HSMM)is introduced and it is used in the proposed work for detecting outliers in the large dataset. The hidden Semi Markov model is widely used in various domains for applications such as MRI sequence analysis, Internet traffic modeling, speech synthesis, anomaly detection, financial time series modeling, etc.,[22].

HSMM is an extension of the Hidden Markov Model (HMM) along with the parameters of HMM, state duration is also considered in HSMM. Generally, the state duration will be an integer value and it is assumed as $d = \{1, 2, 3 \dots s\}$. The major difference between HMM and HSMM is its observations per state. HMM introduces one observation per state whereas HSMM introduces a sequence of observations for each state. Structural comparison between HMM and HSMM is depicted in figure 1 (a) and (b) respectively. The number of observations for the state is obtained from the total amount of time spent in that state.

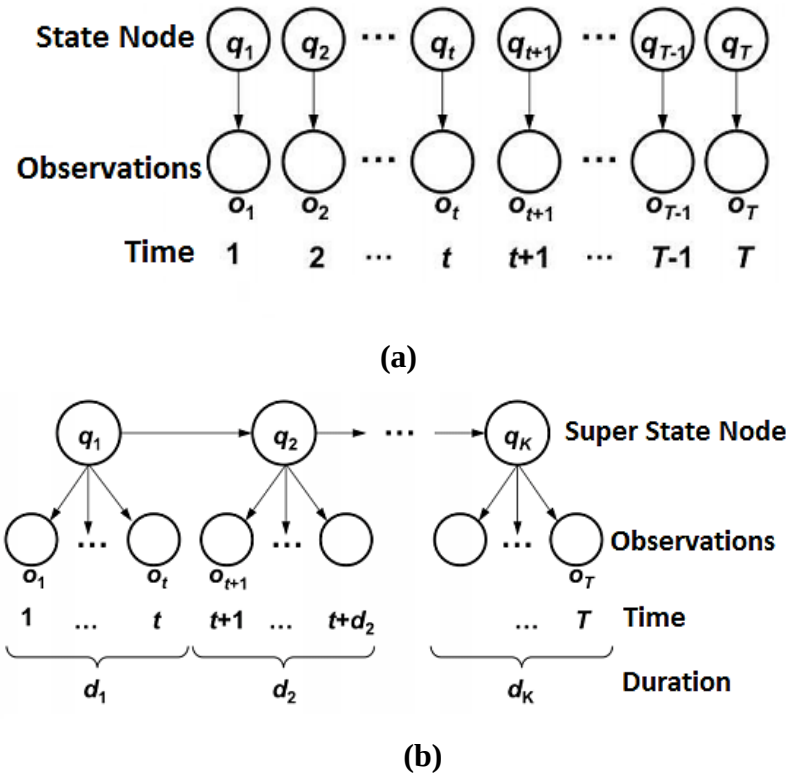

FIGURE 1 Structural comparison (a) HMM (b) HSMM.

For system formulation, let us assume Markov chain in discrete with m hidden states $H(s) = \{1, 2, 3, \dots, m\}$. The sequence of the states are represented as $H(s_{1:t}) = \{s_1, s_2, s_3, \dots, s_t\}$. The observation sequences are denoted as $O(s_{1:t}) = \{o_1, o_2, o_3, \dots, o_t\}$ which is observed over time t and the observed values are represented as $v(s_{1:t}) = \{v_1, v_2, v_3, \dots, v_t\}$. The state transition probability function for one state to another state is given as

$$a_{i,j} = P\left(s_{(t+1:t+m)} = j | s_{(t-m'+1:t)} = i\right) \tag{1}$$

where i, j is the state transition representations that belong to d. It can be observed from equation (1) the transition probability is initiated from $t + 1$ and finished at $t + m$ which clearly indicates that the duration and state depend on the previous state. In

this stage, the observations are emitted and the emission probability is given as

$$b_{j,d} = P\left(o_{(t+1:t+m)} | s_{(t+1:t+m)} = j\right) \tag{2}$$

The above emission probability is independent to time and finally, the initial distribution is given as

$$\pi_{j,d} = P\left(s_{(t-m'+1:t)} = j\right) \tag{3}$$

From equations (1) to (3), the HSMM parameter is defined as

$$\lambda = \left\{a_{i,j}, b_{j,d}, \pi_{j,d}\right\} \tag{4}$$

The transition probability for a state that stayed for a particular duration and it starts to move to another state. This condition is formulated as

$$a_{(i,s)_j} = P\left(s_{(t+1:t+m)} = j | s_{(t-m'+1:t)} = i\right) \tag{5}$$

$$P_j(m) = P\left(s_{(t+1:t+m)} = j | s_{(t-m'+1:t)} = j\right) \tag{6}$$

In the case of independent state transition, the probability is changed as follows

$$a_{(i)_{(s,j)}} = P\left(s_{(t+1:t+m)} = j | s_{(t)} = i\right) \tag{7}$$

The state transition probability for self-transition and independent to the previous state then it is formulated as

$$a_{(i,m')(j,m)} = a_{(i,s)_j} \prod_{\rho=1}^{m-1} a_{jj}(\rho)[1-a_{jj}(m)] \tag{8}$$

where ρ is the state residential time for self-transition. If the state transition is independent to the previous state then $a_{(i,m')(j,m)}$ becomes $a_{jj}P_j(m)$. The forward and backward transitions are given as φ_t and ϑ_t respectively and it is expressed as

$$\varphi_{t(j,m)} = P\left(s_{(t-m'+1:t)} = j, o_{1:t} | \lambda\right) \tag{9}$$

$$\vartheta_{t(j,m)} = P\left(o_{(t+1:t)} | s_{(t-m'+1:t)} = j, \lambda\right) \tag{10}$$

To improve the performance of the HSMM model the parameters (a, b, π, s) are adjusted and presented as enhanced Hidden Semi Markov Model. Generally the coefficients of φ_t and ϑ_t will be in the range of [0,1] and it might affect the computation if the data attributes are very large. In order to solve this, scalar functions are used in φ_t and ϑ_t to scale each state coefficient. The scalars are multiplied. The modified forward transition factors are given as

$$\varphi_i^m(i) = \widehat{\varphi}_i^m(i) / \prod_{s=1}^{t} l^m(s) \tag{11}$$

where $l^m(s)$ is the scalar function and it is given as

$$l^m(s) = 1 / \sum_{\rho=1}^{m} \widehat{\widehat{\varphi}}_i^m(i) \text{ and} \tag{12}$$

$$\widehat{\widehat{\varphi}}_i^m(i) = \sum_{j=1}^{m} \widehat{\varphi}_i^m(i) a_{i,j}, b_{j,d}(o_t) \tag{13}$$

$$\widehat{\varphi}_i^m(i) = \frac{\sum_{j=1}^{m} \widehat{\varphi}_i^m(i) a_{i,j}, b_{j,d}(o_t)}{\sum_{i=1}^{m} \sum_{j=1}^{m} \widehat{\varphi}_i^m(i) a_{i,j}, b_{j,d}(o_t)} \tag{14}$$

Similarly, the backward transition is modified into

$$\vartheta_i^m(i) = \sum_{j=1}^{m} \hat{\vartheta}_{i+1}^m(j) a_{i,j}, b_{j,d}(o_{t+1}) l^m(s) \tag{15}$$

$$\text{where } \hat{\vartheta}_{i+1}^m(j) = l^m(s) \vartheta_i^m(i) \tag{16}$$

$$\hat{\vartheta}_i^m(i) = \frac{\sum_{j=1}^{m} \hat{\vartheta}_i^m(i) a_{i,j}, b_{j,d}(o_{t+1})}{\sum_{i=1}^{m} \sum_{j=1}^{m} \hat{\vartheta}_i^m(i) a_{i,j}, b_{j,d}(o_{t+1})} \tag{17}$$

The summarized pseudocode for the proposed outlier detection model is given as follows

Algorithm 1 Training and detection using Enhanced Hidden Semi Markov Model

Input: Train data $O(s_{1:t})$, Test data $O^*(s_{1:t})$
Process: Initialize Training phase
for $d = \{1, 2, 3 \dots s\}$ **do**
  Assign random values to $a_{i,j}, b_{j,d}, \pi_{j,d}$
  for t=1 to m **do**
    Calculate $a_{i,j}$, $b_{j,d}$, using Eqn.(1) and (2)
    Update HSMM parameter λ Eqn.(1), (2), and (3)
    Update parameters using equation (13) to (17)
    Calculate $\varphi_i^m(i)$ and $\vartheta_i^m(i)$ using Eqn. (11) and (15)
  end for
end for
End process
Process: Initialize detection
for $d = \{1, 2, 3 \dots s\}$ **do**
  for t=1 to m **do**
    Prepare λ from the results obtained in the training phase
    Calculate $\varphi_i^m(i)$ and $\vartheta_i^m(i)$
  end for
end for
End process

4 | RESULT AND DISCUSSION

The performance of proposed outlier detection is demonstrated in this section. parameters such as detection accuracy and detection rate are the major factors considered in this research work and it is observed for three different datasets. MATLAB 14.1 is used for implementation which is installed in an Intel I3 processor with 8GB RAM. Univariate and multivariate datasets are used in the experimentation process. PenDigits[23] and MNIST[24] are univariate datasets in which PenDigits has 6870 samples with 2.27% outlier and MNIST has 7603 samples with 9.2% outliers in the data. For multivariate Wine Quality[25] dataset is utilized and this is one of the standard datasets which is used in various research works. It has 7 classes of data with different percent of outliers. The detailed description for the univariate and multivariate dataset is listed in Table 1 and table 2 .

TABLE 1 Data Description (Single Outlier).

Datasets	Number of Features	Number of samples	Outlier (%)
PenDigits	16	6870	2.27
MNIST	100	7603	9.2

TABLE 2 Data Description (Multiple Outlier).

Datasets	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7
Wine quality	0.41 %	3.32%	29.74%	44.9%	17.96%	3.57%	0.1%

The parameters such as false negative rate, true negative rate are used to obtain the detection rate and detection accuracy for the proposed outlier detection model and it is given in the following equations

$$False\ Negative\ Rate\ (FNR) = \frac{False\ Negatives}{True\ Positives + False\ Negatives} \tag{18}$$

$$True\ Negative\ Rate\ (TNR) = \frac{True\ Negatives}{False\ Positives + True\ Negatives} \tag{19}$$

The detection rate based on equation (17) and (19) is given as

$$Detection\ rate\ (DR) = (1 - FNR) \times TNR \tag{20}$$

$$Detection\ Acc = (DR)/O_{actual} \times 100 \tag{21}$$

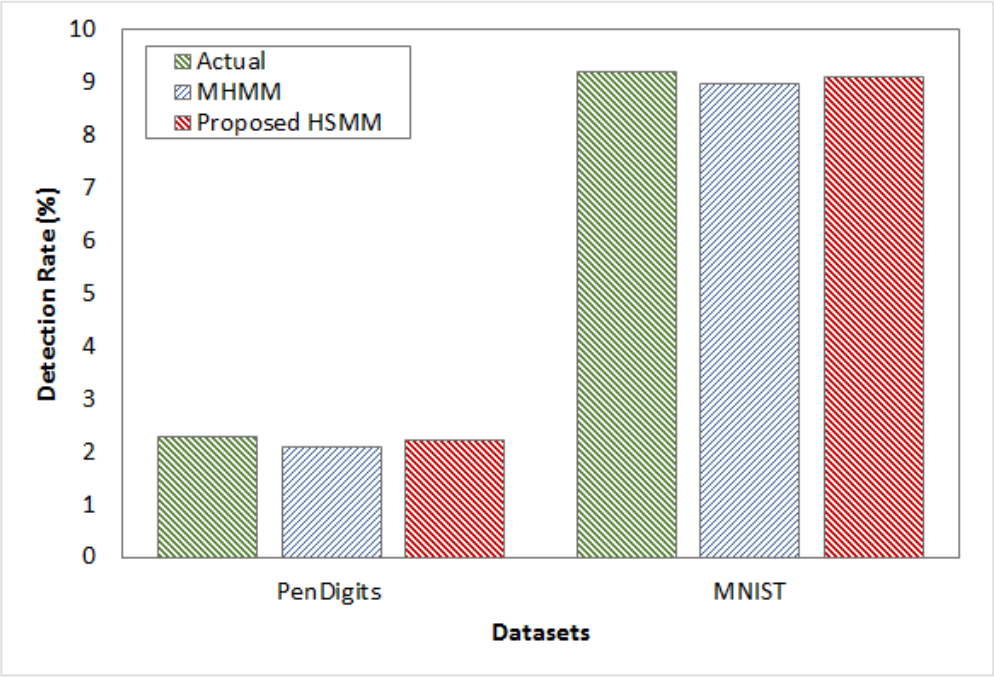

FIGURE 2 Detection Rate (Single Outlier).

The detection rate for the proposed outlier detection model and Modified Hidden Markov Model (MHMM) for outlier detection and actual outliers are compared and depicted in figure 2 . It can be observed from the results, the proposed Hidden Semi Markov Model (HSMM) obtains better results than the MHMM model for both datasets. The difference between actual outlier and detected outlier for PenDigits data set are 6% and 10% for the MNIST dataset for the proposed HSMM model. Whereas the difference for MHMM is 17% for PenDigits dataset and 25% for MNIST dataset which indicates the less performance of MHMM model.

Similarly, the detection rate comparison for the proposed HSMM model and the Modified HMM model for multivariate dataset is depicted in figure 3 . It is observed from the results, proposed HSMM model attains better detection rate than the modified Hidden markov model for all the classes. For few classes, the results are same for both algorithms as it has minimum variations in its data. for class 3 to class 5 maximum detection rate is obtained as 29.21%, 44.6%, 17.5% respectively.

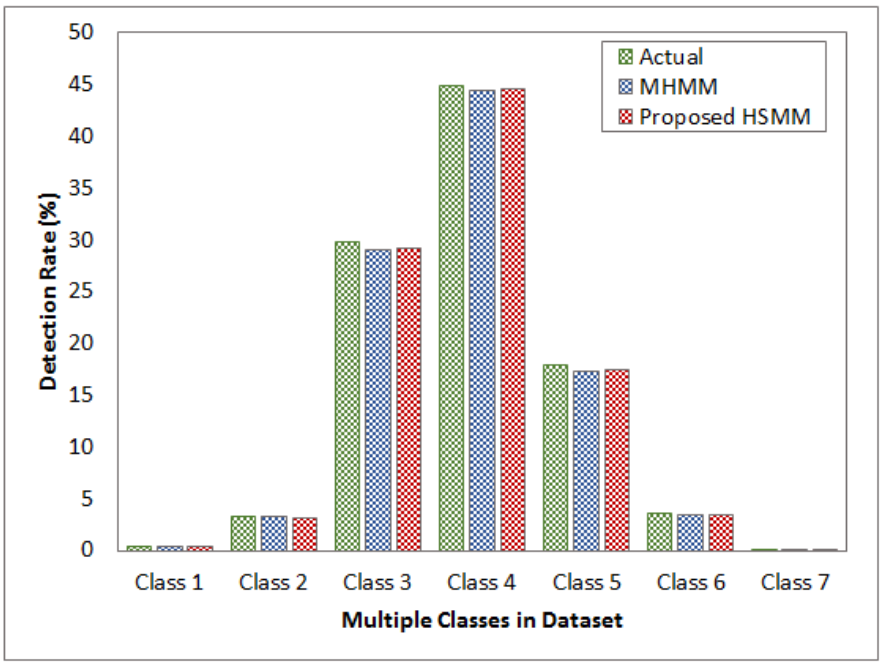

FIGURE 3 Detection Rate (Multiple Outlier).

From the detection rate, the detection accuracy of the proposed model is evaluated for all the three datasets. Figure 5 depicts the comparative analysis of detection accuracy for proposed Hidden semi Markov model (HSMM) and Modified Hidden Markov Model (MHMM). Results depicts that proposed model attains maximum efficiency compared to existing technique due to its efficient detection process.

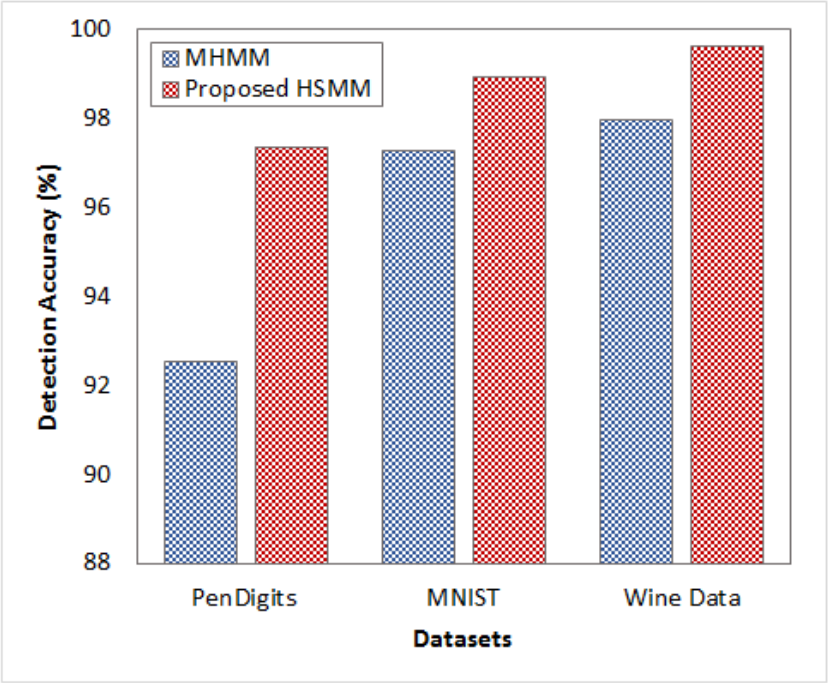

FIGURE 4 Accuracy comparison.

The detection accuracy for PenDigits dataset is approximately 97.3% which is 5% greater than the MHMM model whereas for MNIST dataset the detection accuracy is calculated into 98.8% which is 2% greater than the modified Hidden Markov Model. In case of Multivariate data, the detection accuracy is 99.2% which is 2% greater than existing technique. Overall the proposed Hidden Semi Markov model based outlier detection process obtain an average detection accuracy of 98.62% which is significantly better than the existing Hidden Markov Model-based outlier detection.

5 | CONCLUSION

Outlier detection in the multivariate dataset is essential to improve the performance of data mining models. An enhanced Hidden Semi Markov Model-based outlier detection process is presented in this research work. The limitations in the traditional Hidden Markov model are overcome by the proposed approach in the outlier detection process. The proposed model is experimentally verified and compared with the modified Hidden Markov Model in terms of detection rate and detection accuracy. Results demonstrate that the proposed model has better performance over the existing technique and obtains an average detection accuracy of 98.62%. Proposed research model can be utilized for large dataset with high dimensionality data. The statistical inference for Hidden Semi Markov Model requires more resources which is the limitation of research work. In the future, the research work can be improved by analyzing data depth using kernel functions for better representation of outliers.

References

1. Biao Wang, Zhizhong Mao (2018), “Outlier detection based on Gaussian process with application to industrial processes” Applied Soft Computing, vol. 76, pp. 505–516.

2. Shubin Su; Limin Xiao; Li Ruan; FeiGu; Shupan Li; Zhaokai Wang; Rongbin Xu (2019), “An Efficient Density-Based Local Outlier Detection Approach for Scattered Data” IEEE Access, vol. 7, pp. 1006–1020.

3. PeruriVenkataAnusha, Ch.Anuradha, Patnala S.R. Chandra Murty, Ch. Surya Kiran (2019), “Detecting Outliers in High Dimensional Data Sets Using Z-Score Methodology” International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 1, pp. 1–6.

4. Weiren Yu; Zhengming Ding; Chunming Hu; Hongfu Liu (2019), “Knowledge Reused Outlier Detection” IEEE Access, vol. 7, pp. 43763–43772.

5. Yongmou Li; Yijie Wang; Xingkong Ma; Cheng Qian; Xiaoyong Li (2019), “A Graph-Based Method for Active Outlier Detection with Limited Expert Feedback” IEEE Access, vol. 7, pp. 152267–152277.

6. Biao Wang, Zhizhong Mao, Keke Huang (2019), “Detecting outliers for complex nonlinear systems with dynamic ensemble learning” Chaos, Solitons & Fractals, vol. 121, pp. 98–107.

7. Hongzhi Wang; Mohamed Jaward Bah; Mohamed Hammad (2019), “Progress in Outlier Detection Techniques: A Survey," in IEEE Access, vol. 7, pp. 107964–108000.

8. Li Cheng, Yijie Wang, Xingkong Ma (2019), “A Neural Probabilistic outlier detection method for categorical data” Neurocomputing, vol. 365, pp. 325–335.

9. Anindita Borah, BhabeshNath (2019), “Incremental rare pattern based approach for identifying outliers in medical data” Applied Soft Computing, vol. 85, pp. 1–54.

10. Chao Wang, Zhen Liu, Yan Fu (2019), “VOS: A new outlier detection model using virtual graph Knowledge-Based Systems, vol. 185, pp. 1–12.

11. Chao Wang; Hui Gao; Zhen Liu; Yan Fu (2018), “A New Outlier Detection Model Using Random Walk on Local Information Graph” IEEE Access, vol. 6, pp. 75531–75544.

12. Ping Yang; Dan Wang; Zhuojun Wei; Xiaolin Du; Tong Li (2019), "An Outlier Detection Approach Based on Improved Self-Organizing Feature Map Clustering Algorithm" IEEE Access, vol. 7, pp. 115914–115925.

13. Junli Li; Jifu Zhang; Ning Pang; Xiao Qin (2020), "Weighted Outlier Detection of High-Dimensional Categorical Data Using Feature Grouping" IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 50, no. 11, pp. 4295–4308.

14. Junli Li, Jifu Zhang, YalingXun (2019), "Feature grouping-based parallel outlier mining of categorical data using spark" Information Sciences, vol. 504, pp. 1–19.

15. Kui Yu; Huanhuan Chen (2019), "Markov Boundary-Based Outlier Mining" IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 4, pp. 1259–1264.

16. Jayanta K Dutta, Bonny Banerjee (2019), "Improved outlier detection using sparse coding-based methods" Pattern Recognition Letters, vol. 122, pp. 99–105.

17. Hui Lu; Yaxian Liu; ZongmingFei; Chongchong Guan (2018), "An Outlier Detection Algorithm Based on Cross-Correlation Analysis for Time Series Dataset" IEEE Access, vol. 6, pp. 53593–53610.

18. Leonid Blouvshtein; Daniel Cohen-Or (2021), "Outlier Detection for Robust Multi-Dimensional Scaling" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 9, pp. 2273–2279.

19. Anh Hoang; Toan Nguyen Mau; Duc-Vinh Vo; Van-Nam Huynh (2021), "A Mass-Based Approach for Local Outlier Detection" IEEE Access, vol. 9, pp. 16448–16466.

20. PrabhaVerma; Mousumi Sinha; Siddhartha Panda (2021), "Fuzzy c-Means Clustering-Based Novel Threshold Criteria for Outlier Detection in Electronic Nose" IEEE Sensors Journal, vol. 21, no. 2, pp. 1975–1981.

21. LizhongJin; Junjie Chen; Xiaobo Zhang (2019), "An Outlier Fuzzy Detection Method Using Fuzzy Set Theory," in IEEE Access, vol. 7, pp. 59321–59332.

22. BhavyaMor, SunitaGarhwal, Ajay Kumar (2019), "A Systematic Review of Hidden Markov Models and Their Applications" Archives of Computational Methods in Engineering, pp. 1–20.

23. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository, [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

24. Charu C. Aggarwal and SaketSathe (2015), "Theoretical Foundations and Algorithms for Outlier Ensembles" SIGKDD Explor. Newsl. Vol. 17, no. 1 pp. 24–47.

25. Cortez. P, A. Cerdeira, F. Almeida, T. Matos and J. Reis (2009), "Modeling wine preferences by data mining from physicochemical properties" Decision Support Systems, Elsevier, vol. 47, no. 4, pp. 547–553, 2009.