|  |  |
| --- | --- |
|  |  |
| Unusual Pattern Recognition using Unsupervised learning |  |
|  |  |
|  | 3rd Aug’ 23Internship DRDO |
|  | Kapila Shobit |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | INSTRUCTIONSFor unsupervised learning implementations focused on unusual/anomaly pattern recognition, several algorithms can be used. These algorithms aim to detect anomalies or outliers in data without the need for labeled examples. Some popular algorithms for anomaly detection are:Isolation Forest: This algorithm constructs a random forest to isolate anomalies by recursively partitioning data points. Anomalies are expected to be isolated faster than normal points, making them easier to detect.Local Outlier Factor (LOF): LOF measures the local density deviation of a data point compared to its neighbors. Anomalies are expected to have lower local density compared to their neighbors.One-Class SVM: This algorithm learns a decision boundary around most data points to identify anomalies as points lying outside this boundary.Autoencoders: Autoencoders are neural network architectures used for dimensionality reduction and data reconstruction. Anomalies often have higher reconstruction errors, making them stand out.DBSCAN: Density-Based Spatial Clustering of Applications with Noise is an algorithm that groups dense data points into clusters, considering sparse regions as outliers.Gaussian Mixture Model (GMM): GMM can be used for clustering, but anomalies may be detected as data points with low probability under the fitted Gaussian mixture model.Local Outlier Probability (LoOP): This algorithm combines LOF with probability estimation to give a more reliable anomaly score. | |  |
|  | Process of making an Unsupervised machine learning model: -  Data Collection: Gather data related to the activities you want to monitor. This data can include various features or attributes that describe the activities.  Data Preprocessing: Clean the data and handle any missing values or outliers. Ensure that the data is in a suitable format for the chosen algorithm.  Feature Engineering: Select relevant features or engineer new ones that can help in detecting unusual activities effectively. This step might involve dimensionality reduction techniques to reduce the complexity of the data.  Model Selection: Choose an appropriate unsupervised learning algorithm that suits your use case. As discussed earlier, popular algorithms for anomaly detection include Isolation Forest, Local Outlier Factor, One-Class SVM, etc.  Model Training: Feed the preprocessed data into the selected algorithm and train the model. In the case of some algorithms, you may not need explicit training as they are more like data-driven methods.  Anomaly Detection: After the model is trained, use it to predict anomalies in new, unseen data. Anomalies are data points that deviate significantly from the normal patterns captured by the model.  Threshold Selection: Decide on an appropriate threshold to classify data points as normal or anomalous. This can be done through manual tuning or using techniques like cross-validation to optimize the threshold.  Evaluation: Evaluate the model's performance using appropriate metrics such as precision, recall, F1-score, or ROC curves. You can also use domain-specific metrics if available.  Deployment and Monitoring: Implement the model in your system to continuously monitor activities and detect unusual events in real-time. Regularly update and retrain the model as new data becomes available.  The unsupervised machine learning algorithms that can be used: -  Isolation Forest (iForest): This algorithm isolates anomalies by using random forests to recursively partition data points. Anomalies are expected to be isolated in fewer splits compared to normal data points.  Local Outlier Factor (LOF): LOF measures the local density deviation of a data point with respect to its neighbors. Anomalies have lower local density compared to their neighbors.  One-Class Support Vector Machines (One-Class SVM): This algorithm learns a decision boundary around most data points in a high-dimensional space, identifying anomalies as points outside this boundary.  Autoencoders: Autoencoders are neural networks used for unsupervised representation learning and data reconstruction. Anomalies result in higher reconstruction errors, making them stand out.  Density-Based Spatial Clustering of Applications with Noise (DBSCAN): DBSCAN groups dense data points into clusters and marks points in low-density regions as outliers.  Gaussian Mixture Model (GMM): GMM is primarily used for clustering, but it can be adapted to detect anomalies by identifying data points with low probability under the fitted Gaussian model.  K-Nearest Neighbors (KNN): In unsupervised anomaly detection, KNN can be used to measure the distance of a data point to its k-nearest neighbors, and anomalies have larger distances.  Histogram-based Outlier Score (HBOS): HBOS creates histograms for each feature and computes an outlier score based on the joint probability of feature occurrences.  Cluster-Based Local Outlier Factor (CBLOF): CBLOF first clusters the data and then uses LOF to compute the anomaly score based on the distance to the cluster centers.  Angle-Based Outlier Detection (ABOD): ABOD measures the angles between data points and uses them to compute an outlier score. |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |
|  |