International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 03 | March - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Use of Machine Learning Classifiers on Structured Vs Unstructured Data

Dr. Syed Tabrez Hassan[1], Asha Parvin[2], Aihik Datta[2], Sayantan Mondal[2]

[1]Associate Professor,

[2]Student of BBA-Business Analytics

Adamas University, Kolkata

Abstract

This study investigates the efficacy of machine learning classifiers on structured versus unstructured data. Structured data, organized in predefined formats, enables conventional analysis. Conversely, unstructured data, lacking predefined structures, presents challenges in processing. We assess classifier performance across diverse datasets, focusing on accuracy, efficiency, and adaptability. Results reveal distinct classifier behaviors: structured data favors traditional algorithms, while unstructured data necessitates advanced techniques such as natural language processing and deep learning. Understanding these nuances enhances data-driven decision-making in various domains, from business intelligence to sentiment analysis and image recognition.

Key words: Structured Data, Unstructured Data, Machine Learning, Classifiers

Introduction:

Structured data refers to information that is organized in a predefined format with well-defined fields and categories. This type of data is typically found in relational databases or spreadsheets, where each data element is labeled and stored in a specific location. Structured data is easy to analyze, search, and process using traditional database management systems. Examples include tables of customer information with fields like name, address, and phone number, or sales data organized by product, date, and quantity.

On the other hand, unstructured data refers to information that does not have a predefined data model or is not organized in a pre-defined manner. This type of data often includes text-heavy content, such as emails, social media posts, images, videos, and documents. Unstructured data lacks a clear structure or format, making it more challenging to analyze and process using traditional methods. However, advancements in

natural language processing (NLP), machine learning, and other technologies have enabled organizations to extract valuable insights from unstructured data, unlocking its potential for various applications like sentiment analysis, content categorization, and image recognition.

Structured Vs Unstructured Data

Feature	Structured Data	Unstructured Data
Format	Organized in tables or databases	Lack of predefined structure
Storage	Efficient and easily searchable	May require more sophisticated storage solutions
Representation	Clear and well-defined schema	Varied formats, such as text, images, audio, and video
Examples	Spreadsheets, relational databases, CSV files	Text documents, emails, social media posts, multimedia content
Use Cases	Financial records, inventory management, customer databases	Social media analysis, sentiment analysis, image recognition, natural language processing
Flexibility	Rigid structure with predefined relationships	Flexible, capturing diverse and unanticipated information
Accessibility	Easily query able and accessible	More challenging to query directly, may need specialized tools or algorithms
Storage and Retrieval	Efficient due to organized structure	Retrieval may be complex, especially for large datasets
Complexity	Generally less complex in terms of analysis.	More complex, requiring advanced processing methods

Table 1: Structured vs Unstructured Data

What is a Classifier?

In the context of machine learning, a classifier is a model or algorithm that is trained to categorize input data into predefined classes or categories. It is a type of supervised learning technique where the algorithm learns from labeled training data to make predictions on new, unseen data. Classifiers are used to solve classification problems, where the goal is to assign a label or category to input data based on its features.

Significance of classifiers in Machine Learning

Classifiers hold immense significance in the realm of machine learning, acting as the cornerstone for various tasks and applications. Here's a breakdown of their importance:

1. Organization and Decision Making: Classifiers allow us to categorize data into predefined groups or classes.
2. Make sense of large datasets: By classifying data points, we can identify patterns, trends, and anomalies that would be difficult to discern otherwise.
3. Simplify complex information: Classifiers enable us to condense information into manageable categories, facilitating informed decision-making.
4. Automate tasks: Classification algorithms can automate tasks like spam filtering in emails, categorizing customer reviews, or flagging fraudulent transactions. Prediction and Forecasting. Classifiers can learn from past data to predict the class label of new, unseen data points.
5. Risk assessment: Classifiers can assess the risk of loan defaults, credit card fraud, or medical complications based on historical data.
6. The Foundation for Other Techniques: Classifiers serve as the building blocks for more complex machine learning techniques.
7. Ensemble methods: These techniques combine the predictions of multiple classifiers to improve overall accuracy and robustness.
8. Deep learning: Many deep learning architectures, such as convolutional neural networks, are essentially sophisticated classifiers for various data types like images, text, or audio.

Classifiers are fundamental to machine learning, enabling data organization, and prediction, and forming the foundation for more advanced techniques. Their ability to categorize, analyze, and make sense of information propels various applications across diverse fields, making them an invaluable tool in the machine-learning landscape.

Types of Classifiers used in Machine Learning:

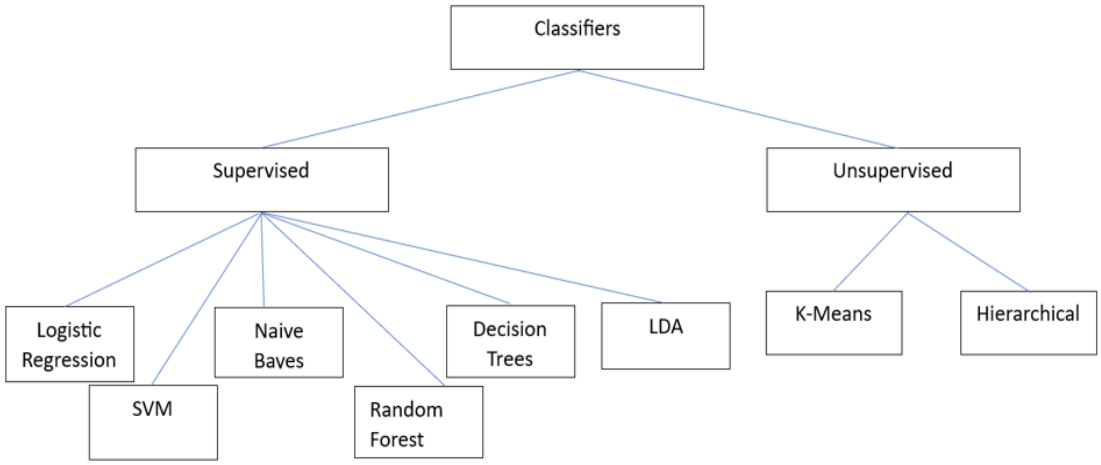

Figure 1: Types of Classifiers

Supervised Classifiers:

1. **Decision Trees**: Decision trees are hierarchical structures where each internal node represents a decision based on a feature, and each leaf node represents a class label.
    - Key Characteristics: Easy to interpret and visualize, can handle both numerical and categorical data, Prone to overfitting, especially with complex trees.
2. **Support Vector Machines (SVM):** SVM is a discriminative classifier that finds the hyperplane that best separates classes in feature space.
    - Key Characteristics: Effective in high-dimensional spaces, Versatile due to various kernel functions, Memory efficient since it only uses a subset of training points for decision making.
3. **Logistic Regression**: Logistic regression models the probability of a binary outcome using a logistic function.
    - Key Characteristics: Provides probabilistic interpretations of outcomes, can handle linear and nonlinear relationships, sensitive to outliers.

4. **Random Forest:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

    - Key Characteristics: Combines multiple decision trees to reduce overfitting, handles high-dimensional datasets well, provides feature importance estimation, can handle missing values and maintain accuracy with unbalanced data.

5. **Naive Bayes**: Naive Bayes classifiers are probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features.

    - Key Characteristics: Simple and fast to train, works well with high-dimensional data, robust to irrelevant features, assumes independence between features, which might not hold true in all cases.

6. **Linear Discriminant Analysis (LDA):** LDA is a statistical technique used for dimensionality reduction and classification. It seeks to find a linear combination of features that best separates two or more classes of objects. LDA maximizes the separation between classes while minimizing the variance within each class.

    - Key Characteristics: Dimensionality reduction and classification technique, seeks linear combination of features for class separation, maximizes inter-class separation while minimizing intra-class variance, often used in pattern recognition, machine learning, and statistics

Unsupervised Classifiers:

1. K-Means Clustering: K-Means is a partitioning method that partitions data into k clusters where each data point belongs to the cluster with the nearest mean.

    - Key Characteristics: Simple and computationally efficient, requires specifying the number of clusters (k) beforehand, sensitive to initialization and outliers.

2. Hierarchical Clustering: Hierarchical clustering builds a tree of clusters, where the leaf nodes are individual data points, and the internal nodes represent clusters of varying granularity.

- Key Characteristics: No need to specify number of clusters beforehand, Produces dendrograms for visual interpretation, computationally expensive for large datasets.

Table 2: Applied examples of Classifiers on Structured data

Classifiers Suitable for Structured Data	Examples
1. Decision Trees:	- Example: Predicting customer purchases based on demographic data.
2. Support Vector Machines (SVM):	- Example: Classifying emails as spam or not spam based on features.
3. k-Nearest Neighbors (k-NN):	- Example: Predicting movie genres based on ratings and duration.
4. Naive Bayes:	- Example: Categorizing news articles into topics based on word frequency.
5. Logistic Regression:	- Example: Predicting loan approval based on financial structured data.

Table 3: Applied examples of Classifiers on Unstructured data

Classifiers Suitable for Unstructured Data	Examples
1. Neural Networks:	- Example: Image classification for identifying objects in photographs.
2. Support Vector Machines (SVM):	- Example: Recognizing handwriting patterns for digit recognition.
3. Random Forest:	- Example: Analyzing unstructured text data to predict sentiment.
4. Nearest Neighbors (k-NN):	- Example: Face recognition based on patterns in unstructured image data.
5. Naive Bayes:	- Example: Classifying emails as spam or not spam based on text content.
6. Deep Learning Models (e.g., CNNs):	- Example: Natural language processing tasks like sentiment analysis.

Conclusion:

In conclusion, the comparative analysis of machine learning classifiers on structured and unstructured data underscores the importance of tailored approaches for distinct data types. While structured data benefits from conventional classifiers due to its organized format, unstructured data demands advanced techniques like natural language processing and deep learning. Understanding the nuances between structured and unstructured data enhances decision-making across diverse domains, unlocking valuable insights from both traditional databases and text-heavy sources like social media and documents. Further research in this area will facilitate the development of robust methodologies for extracting knowledge from the ever-expanding volume of structured and unstructured data.

References:

Shi, Y., Sagduyu, Y., & Grushin, A. (2017, April). How to steal a machine learning classifier with deep learning. In *2017 IEEE International symposium on technologies for homeland security (HST)* (pp. 1-5). IEEE.

Zhang, J., Wang, Z., & Verma, N. (2016, June). A machine-learning classifier implemented in a standard 6T SRAM array. In *2016 ieee symposium on vlsi circuits (vlsi-circuits)* (pp. 1-2). IEEE.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, *45*(1), S199-S209.

Wang, L. (2019, December). Research and implementation of machine learning classifier based on KNN. In *IOP Conference Series: Materials Science and Engineering* (Vol. 677, No. 5, p. 052038). IOP publishing.

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3* (pp. 246-252). Springer Berlin Heidelberg.

Kulkarni, V. Y., & Sinha, P. K. (2013). Random forest classifiers: a survey and future research directions. *Int. J. Adv. Comput*, *36*(1), 1144-1153.

Jizat, J. A. M., Majeed, A. P. A., Nasir, A. F. A., Taha, Z., & Yuen, E. (2021). Evaluation of the machine learning classifier in wafer defects classification. *ICT Express*, *7*(4), 535-539.

Nguyen, P. T., Di Rocco, J., Iovino, L., Di Ruscio, D., & Pierantonio, A. (2021). Evaluation of a machine learning classifier for metamodels. *Software and Systems Modeling*, *20*(6), 1797-1821.

Narudin, F. A., Feizollah, A., Anuar, N. B., & Gani, A. (2016). Evaluation of machine learning classifiers for mobile malware detection. *Soft Computing*, *20*, 343-357.

Melville, P., & Mooney, R. J. (2003, August). Constructing diverse classifier ensembles using artificial training examples. In *Ijcai* (Vol. 3, pp. 505-510).

Kratsch, W., Manderscheid, J., Röglinger, M., & Seyfried, J. (2021). Machine learning in business process monitoring: a comparison of deep learning and classical approaches used for outcome prediction. *Business & Information Systems Engineering*, *63*, 261-276.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47.

Majnik, M., & Bosnić, Z. (2013). ROC analysis of classifiers in machine learning: A survey. *Intelligent data analysis*, *17*(3), 531-558.

Soni, M., Chauhan, S., Bajpai, B., & Puri, T. (2020, September). An approach to enhance fall detection using machine learning classifier. In *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 229-233). IEEE.