Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2020, 7(3):31-36

Research Article ISSN: 2394 - 658X

Data Classification and Detection Using Machine Learning

Khirod Chandra Panda

Principal Engineer | Asurion
khirodpanda4bank@gmail.com

ABSTRACT

The passage discusses the increasing use of the internet in modern society and its impact on various aspects of life, particularly in the realm of e-commerce and online transactions. It highlights the convenience and flexibility that online shopping and bill payment offer, attributing this to advancements in technology such as online banking and credit card payments. However, with the rise in online transactions comes a corresponding increase in credit card fraud, posing challenges for banks in distinguishing between legitimate and fraudulent transactions, especially in cases where customers lose their credit cards. To address this issue, the paper proposes the use of Machine Learning (ML) in real-time to detect and prevent fraud transactions. It emphasizes the importance of using an unsupervised approach to detect anomalies, which can lead to higher accuracy in identifying fraudulent activities. The proposed ML-based anomaly detection process is described as having broad applicability, including fraud detection in banking, billing discrepancies in telecommunications, security monitoring, network traffic analysis, healthcare, and various manufacturing industries. Furthermore, the paper explains that unsupervised learning can also serve as a form of dimension reduction, particularly useful when dealing with datasets with numerous features but limited cases. By condensing the features into a smaller set of dimensions while retaining most of the original information, unsupervised learning can help analysts gain insights into the underlying structure of the data. For example, it can reveal geographic patterns or common features among geographically distant locations, providing valuable insights for decision-making in various industries.

Key words: Data Classification, Machine Learning

INTRODUCTION

Credit card fraud is a significant issue in the online sector, involving the unauthorized use of someone else's credit card for payments or cash withdrawals. Victims of credit card fraud may not realize they are affected until substantial damage has been done to their credit, often taking years to recover. Thus, prevention is preferable to cure in this scenario. Many machine learning algorithms perform best when the number of instances in each class is roughly equal. Imbalanced datasets, where one class significantly outweighs the other, pose challenges.

To address credit card fraud detection, various techniques are employed, including genetic algorithms, artificial neural networks, regression, decision trees, and random forests. However, credit card transaction datasets are often scarce, highly imbalanced, and skewed. Selecting optimal features for these models is crucial in evaluating their performance. Imbalanced datasets can affect the learning phase and prediction of machine learning algorithms. Ensemble methods are used to reduce variance in the dataset, providing effective and accurate results. Machine learning is a field focused on computer algorithms that learn and perform specific tasks without explicit programming. It enables computers to learn from past experiences and make better decisions in the future. Arthur Samuel described “machine learning as giving computers the ability to learn without being explicitly programmed”. It involves discovering algorithm structures that facilitate learning from data. Applications of machine learning include sentiment analysis, email spam detection, targeted advertisements, recommendation engines, and pattern mining for market basket analysis.

Sometimes, machine learning and Artificial Intelligence (AI) are used interchangeably. However, machine learning and AI are two distinctive areas of computing.

Machine learning [1] can be categorized into three types: supervised learning, where the algorithm learns from labeled data; unsupervised learning, where the algorithm learns from unlabeled data; and reinforcement learning, where the algorithm learns by interacting with its environment and receiving rewards or penalties.

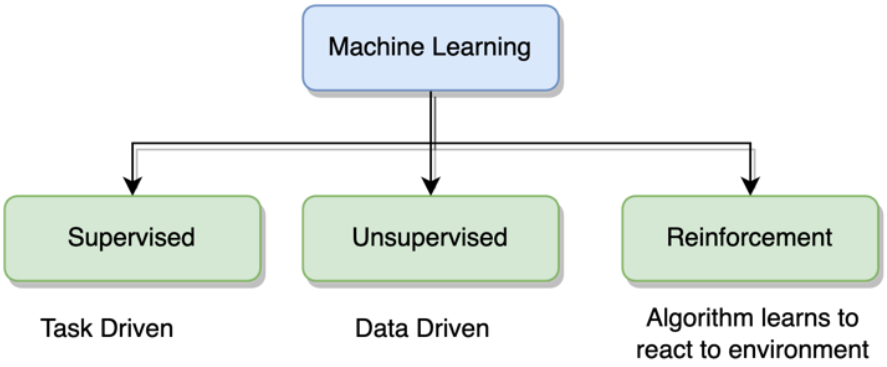


Fig. 1 Types of Machine Learning

LITERATURE REVIEW

Supervised Learning

Supervised learning algorithms are a subset of machine learning techniques where the model learns patterns from labeled data, which includes both input features and corresponding output labels. The primary goal of supervised learning is to learn a mapping function from input variables to output variables.

The learning process involves presenting the algorithm with a training dataset consisting of input-output pairs. The algorithm then learns the relationship between the input and output by adjusting its internal parameters, such as weights and biases, to minimize the error between the predicted output and the actual output. This process is often referred to as training or fitting the model.

Once the model has been trained, it can be used to make predictions on new, unseen data. The model applies the learned mapping function to the new input data to predict the corresponding output. The performance of the model is evaluated based on how well it predicts the output for the new data.

Supervised learning is widely used in various applications, such as classification, regression, and ranking. Classification tasks involve predicting a discrete label or category, while regression tasks involve predicting a continuous value. Ranking tasks involve ordering a set of items based on their relevance or importance.

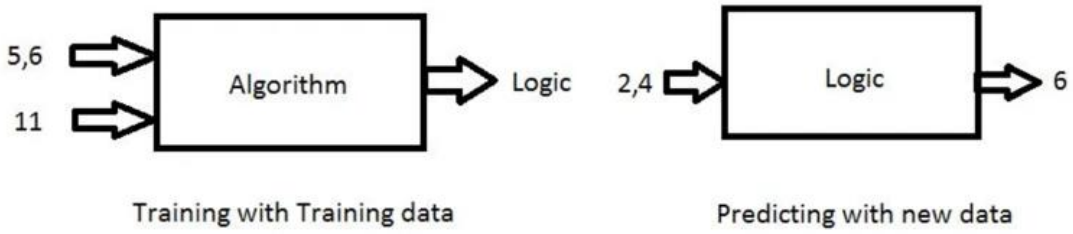
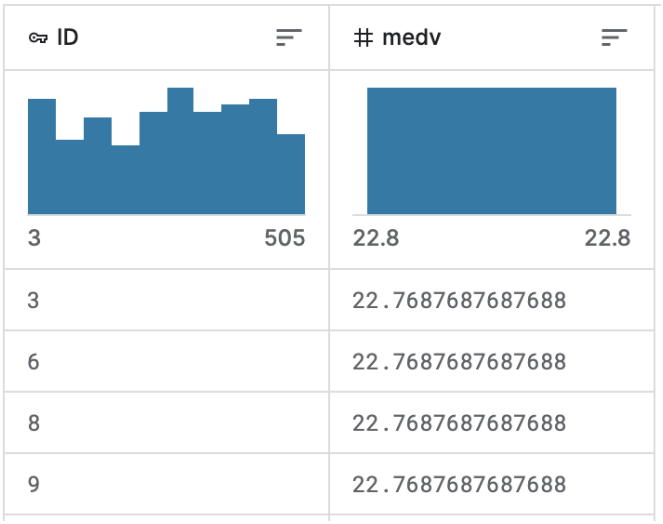


Fig. 2

We first train the model with the lots of training data (inputs & targets), then with new data and the logic we got before we predict the output.

(Note: We don't get exact 6 as answer we may get value which is close to 6 based on training data and algorithm)

In supervised learning, the data that the model is trained on includes a label for each observation where the label is the correct answer. For a numerical dataset like the Boston housing prices dataset [2], the label would be the price of the house.



ID	medv
3 – 505	22.8 – 22.8
3	22.7687687687688
6	22.7687687687688
8	22.7687687687688
9	22.7687687687688

Fig. 3

For a categorical dataset such as the iris dataset [3], the label would be the species of the flower.

Variable Name	Role	Type	Demographic	Description	Units
sepal length	Feature	Continuous			cm
sepal width	Feature	Continuous			cm
petal length	Feature	Continuous			cm
petal width	Feature	Continuous			cm
class	Target	Categorical		class of iris plant: Iris Setosa, Iris Versicolour, or Iris Virginica	

Fig. 4

Supervised learning is generally broken into 2 categories.

- Regression
- Classification

Regression

For regression, the goal is to predict a continuous number. The above referenced Boston housing prices dataset [2] is an example of a regression problem. Fitness provides a plethora of examples of regression problems. If I want to try to predict heart rate, the number of calories burned, or the age of the participant those would all be regression problems.

Classification

For classification, the goal is to predict a category. An example of a classification problem is the above referenced iris dataset. The categories are the type of flower. Another example of classification is the titanic dataset [4]. People either survived the sinking of the ship or they didn't.

An important distinction within classification is binary classification vs multiclass classification.

- Binary classification There are two and only two labels. Example: Titanic dataset. Survived: yes or no.
- Multiclass classification There are more than two labels. Example: iris dataset. Species of flower: Serosa, Versicolor, or Virginica

There is another type of classification, multilabel classification. An example of this would be browser used by a customer for an online retailer. There's nothing stopping the customer from using more than one browser and so the observation for that customer would have more than one label. I haven't worked with multilabel problems, and they don't seem particularly prevalent, so this will probably be as in-depth as I go into that topic.

A situation could arise wherein there are so many categories, and the categories are numeric, that it becomes difficult to tell if the problem should be categorized as regression or multiclass. If the values of the label are continuous or fall along a scale (like centimeters on a ruler) then the problem is a regression problem. If the values are discrete and separate, the problem is a classification problem.

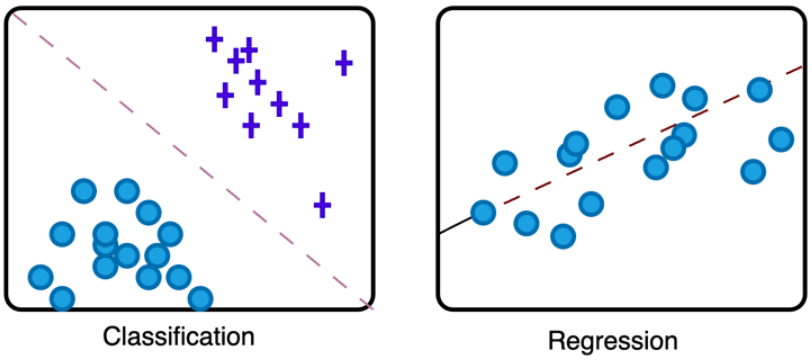


Fig. 5 Classification separates the data, Regression fits the data.

For example, if the label included the ages of children between 5 and 18 years, it may be tempting to think of this as a classification problem. However, because the ages fall on a scale, it should be a regression problem. On the other hand, if we had grouped the children into child, youth, preteen, and teenager then it would be a classification problem.

The purpose of supervised learning is to take information that is known and generalize patterns in that information to predict values for data that has never been seen before.

For this learning style, every observation in the training dataset must have a label. To appropriately evaluate how the model performed, every observation in the testing dataset must also have a label. However, a label isn't required for the express purpose of making a prediction - just determining how good the prediction is.

Unsupervised Learning

The training data does not include Targets here so we don't tell the system where to go, the system has to understand itself from the data we give. Unsupervised Algorithms are the algorithms that are used for unlabeled data set, i.e., which doesn't have the output variable. These algorithms are used to find the unknown patterns in the data and are used to analyze and segment them into clusters based on the behaviors. Similar behavioral patterns of the data are clustered into similar groups. These algorithms are widely useful for the unlabeled dataset [5].

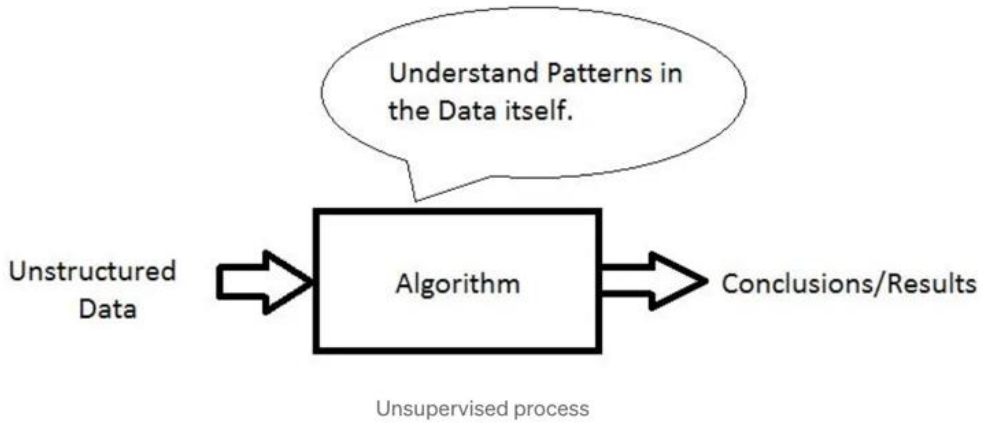


Fig. 6

The algorithm doesn't have any prior knowledge or training data. It just converts it into pixels and groups them based on the data provided. In unsupervised learning, we group the parts of data based on similarities within each other. The data in unsupervised learning is unlabeled, meaning there are no column names. This is not important because we don't have any specific knowledge/training of the data. Unsupervised learning problems can be further grouped as clustering and association problems:

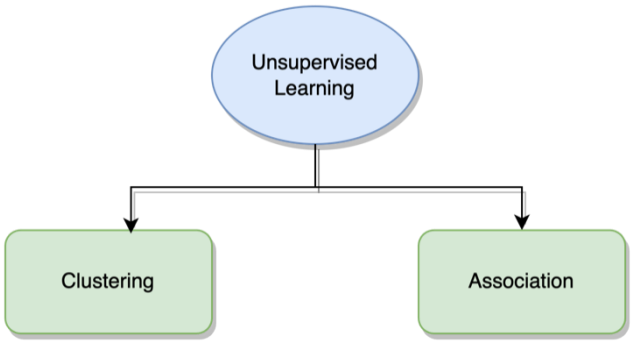


Fig. 7

Clustering: is a technique for grouping objects into clusters based on their similarities, ensuring that objects within a cluster are more like each other than to those in other clusters. Cluster analysis identifies commonalities among data objects and organizes them based on the presence or absence of these commonalities.

Association: is an unsupervised learning approach used to discover relationships between variables in a large dataset. It identifies sets of items that frequently occur together in the dataset. Association rules enhance marketing strategies by revealing patterns such as customers who purchase item X (e.g., bread) are also likely to buy item Y (e.g., butter/jam). A classic example of association rules is Market Basket Analysis.
Unsupervised learning is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
Unsupervised learning is preferable as it is easy to get unlabeled data in comparison to labeled data.
Unsupervised learning is intrinsically more difficult than supervised learning as it does not have corresponding output.
The result of the unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

Reinforcement Learning
Reinforcement learning (RL) is a machine learning paradigm where an agent interacts with an environment, learns to make decisions, and improves its performance over time through trial and error. Unlike supervised learning, RL does not require labeled data; instead, the agent learns from the consequences of its actions, receiving positive feedback for desirable actions and negative feedback for undesirable ones.
The core idea behind RL is to find a strategy (policy) that maximizes the cumulative reward obtained by the agent over time. This is achieved by iteratively exploring the environment, taking actions, and updating the policy based on the observed rewards. The agent's goal is to learn an optimal policy that guides its actions to achieve the best possible outcome in the given environment.
RL has numerous applications, including game playing, robotics, resource management, and autonomous driving. In a game-playing scenario, an RL agent might learn to navigate a maze or play a game like chess or Go by trial and error, gradually improving its performance through experience. In robotics, RL can be used to teach a robot to perform complex tasks such as grasping objects or navigating an environment.
One of the key challenges in RL is the trade-off between exploration and exploitation. The agent must explore the environment to discover potentially beneficial actions, but it also needs to exploit known good actions to maximize its reward. Balancing these two aspects is crucial for efficient learning and optimal decision-making.
Overall, RL is a powerful approach to machine learning that enables agents to learn complex behaviors and make decisions in dynamic and uncertain environments. By combining exploration and exploitation, RL algorithms can adapt to changing conditions and learn to achieve optimal performance over time.

Terms Used in Reinforcement Learning

- Agent (): An entity that can perceive/explore the environment and act upon it.
- Environment (): A situation in which an agent is present or surrounded by. In RL, we assume the stochastic environment, which means it is random in nature.
- Action (): Actions are the moves taken by an agent within the environment.
- State (): State is a situation returned by the environment after each action taken by the agent.
- Reward (): A feedback returned to the agent from the environment to evaluate the action of the agent.
- Policy (): Policy is a strategy applied by the agent for the next action based on the current state.
- Value (): It is expected long-term retuned with the discount factor and opposite to the short-term reward.

- Q-value (): It is mostly like the value, but it takes one additional parameter as a current action.

in the case of reinforcement learning the goal is to find a suitable action model that would maximize the total cumulative reward of the agent. The figure below illustrates the action-reward feedback loop of a generic RL model.

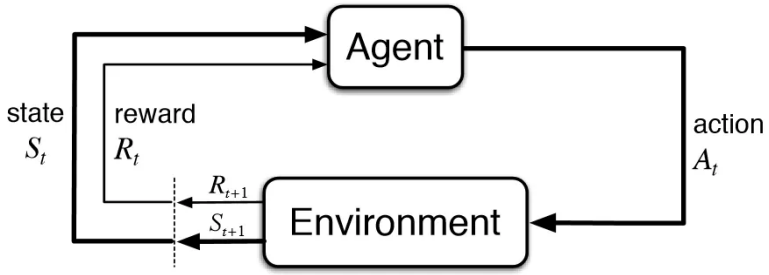


Fig. 8

CONCLUSION

Machine Learning Research spans almost four decades. Much of the research has been to define various types of learning, establish the relationships among them, and elaborate on the algorithms that characterize them [7]. But much less effort has been devoted to bringing machine learning to bear on real-world applications. But recently, researchers have found broader applications of machine learning to real-world problems. Some of these are:

- Bioinformatics
- Brain-machine interfaces
- Classifying DNA sequences
- Computational finance

Computer vision, including object recognition.

REFERENCES

[1]. What is Machine Learning., [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Accessed Feb 2020].

[2]. Boston Housing | Kaggle. [Online] Available: https://www.kaggle.com/competitions/boston-housing/data

[3]. IRIS - UCI Machine Learning Repository. [Online]Available: https://archive.ics.uci.edu/dataset/53/iris

[4]. Titanic - Machine Learning from Disaster | Kaggle. [Online] Available: https://www.kaggle.com/c/titanic/data

[5]. Popular Machine Learning Methods, [Online]. Available: https://www.sas.com/enus/insights/analytics/machinelearning.html. [Accessed Feb 2020].

[6]. Bin Xu; Chenguang Yang; Zhongke Shi., Reinforcement Learning Output Feedback NN Control Using Deterministic Learning Technique., IEEE Transactions on Neural Networks and Learning Systems, 25(3) (2014).

[7]. Alberto Maria Segre., Applications of Machine Learning., Cornell University.