A review on unsupervised machine learning for networking

Education, Learning

Abstract: Recently, there has been a rising trend of employing unsupervised machine learning using unstructured raw network data to improve network performance and provide services, such as traffic engineering, anomaly detection, Internet traffic classification, and quality of service optimization.

We provide an extensive overview highlighting recent advancements in unsupervised learning techniques, and describe their applications in various learning tasks, in the context of networking. The growing interest in applying unsupervised learning techniques in networking stems from their great success in other fields, such as computer vision, natural language processing, speech recognition, and optimal control. The focus of this survey paper is to provide an overview of applications of unsupervised learning in the domain of networking.

In addition, unsupervised learning can unconstraint us from the need for labeled data and manual handcrafted feature engineering, thereby facilitating flexible, general, and automated methods of machine learning. While machine learning andartificial intelligencehave long been applied in networking research, the bulk of such works has focused on supervised learning. Through this timely review, we aim to advance the current state of knowledge, by carefully synthesizing insights from previous survey papers, while providing contemporary coverage of the recent advances and innovations.

Introduction

Networks - such as the Internet and mobile telecom networks - serve the function of the central hub of modern human societies, which the various

threads ofmodern lifeweave around With networks becoming increasingly dynamic, heterogeneous, and complex, the management of such networks has become less amenable to manual administration, and it can benefit from leveraging support from The associate editor coordinating the review of this manuscript and approving it for publication was Nuno Garcia. methods for optimization and automated decision-making from the fields of artificial intelligence (AI) and machine learning (ML).

Such AI and ML techniques have already transformed multiple fields—e. g., computer vision, natural language processing (NLP), speech recognition, and optimal control (e. g., for developing autonomous self-driving vehicles)—with the success of these techniques mainly attributed to firstly, significant advances in unsupervised ML techniques such as deep learning, secondly, the ready availability of large amounts of unstructured raw data amenable to processing by unsupervised learning algorithms, and finally, advances in computing technologies through advances such as cloud computing, graphics processing unit (GPU)technologyand other hardware enhancements.

It is anticipated that AI and ML will also make a similar impact on the networking ecosystem and will help realize a future vision of cognitive networks. In order to develop a reliable data-driven network, data quality must be taken care before subjecting it to an appropriate unsupervised ML.

The purpose of this paper is to highlight the important advances in unsupervised learning, and after providing a tutorial introduction to these

techniques, to review how such techniques have been, or could be, used for various tasks in modern next-generation networks comprising both computer networks as well as mobile telecom networks.

Techniques of Unsupervised Learning

In this section, we will introduce some widely used unsupervised learning techniques and their applications in computer networks.

Hierarchical Learning

Hierarchical learning is defined as learning simple and complex features from a hierarchy of multiple linear and nonlinear activations. Hierarchical learning is intimately related to how deep learning is performed in modern multi-layer neural networks.

In particular, deep learning techniques benefits from the fundamental concept of artificial neural networks (ANNs), a deep structure consists of multiple hidden layers with multiple neurons in each layer, a nonlinear activation function, a cost function, and a back-propagation algorithm. Deep learning is a hierarchical technique that models high level abstraction in data using many layers of linear and nonlinear transformations.

Data Clustering

Clustering is an unsupervised learning task that aims to find hidden patterns in unlabeled input data in the form of clusters. Simply put, it encompasses the arrangement of data in meaningful natural groupings on the basis of the

similarity between different features (as illustrated in Fig. 1) to learn about its structure.

Clustering involves the organization of data in such a way that there are high intra-cluster and low inter-cluster similarity. The resulting structured data is termed as data-concept. . Clustering is used in numerous applications from the fields of ML, data mining, network analysis, pattern recognition, and computer vision.

Clustering improves performance in various applications. McGregor propose an efficient packet tracing approach using the Expectation-Maximization (EM) probabilistic clustering algorithm, which groups flows (packets) into a small number of clusters, where the goal is to analyze network traffic using a set of representative clusters.

Latent Variable Models

A latent variable model is a statistical model that relates the manifest variables with a set of latent or hidden variables. Latent variable model allows us to express relatively complex distributions in terms of tractable joint distributions over an expanded variable space. More formally, a latent variable model (LVM) p is a probability distribution over two sets of variables x, z : p(x, z; ?), where the x variables are observed at learning time in a dataset D and the z are never observed. The model may be either directed or undirected.

Underlying variables of a process are represented in higher dimensional space using a fixed transformation, and stochastic variations are known as latent variable models where the distribution in higher dimension is due to small number of hidden variables acting in a combination. These models are used for data visualization, dimensionality reduction, optimization, distribution learning, blind signal separation and factor analysis.

Dimensionality Reduction

Representing data in fewer dimensions is another well-established task of unsupervised learning. Real world data often have high dimensions—in many datasets, these dimensions can run into thousands, even millions, of potentially correlated dimensions.
However, it is observed that the intrinsic dimensionality (governing parameters) of the data is less than the total number of dimensions.

In order to find the essential pattern of the underlying data by extracting intrinsic dimensions, it is necessary that the real essence is not lost; e. g., it may be the case that a phenomenon is observable only in higher-dimensional data and is suppressed in lower dimensions, these phenomena are said to suffer from the curse of dimensionality. While dimensionality reduction is sometimes used interchangeably with feature selection, a subtle difference exists between the two.

Feature selection is traditionally performed as a supervised task with a domain expert helping in handcrafting a set of critical features of the data. Such an approach generally can perform well but is not scalable and prone

to judgment bias. Dimensionality reduction, on the other hand, is more generally an unsupervised task, where instead of choosing a subset of features, it creates new features (dimensions) as a function of all features. Said differently, feature selection considers supervised data labels, while dimensionality reduction focuses on the data.

Outlier Detection

Outlier detection is an important application of unsupervised learning. A sample point that is distant from other samples is called an outlier. An outlier may occur due to noise, measurement error, heavy tail distributions and a mixture of two distributions.

There are two popular underlying techniques for unsupervised outlier detection upon which many algorithms are designed, namely the nearest neighbor based technique and clustering based method. The nearest neighbor method works on estimating the Euclidean distances or average distance of every sample from all other samples in the dataset. Clustering based methods use the conventional K-means clustering technique to find dense locations in the data and then perform density estimation on those clusters.

From above techniques we can draw below points:

- Hierarchical learning techniques are the most popular schemes in literature for feature detection and extraction.

- Learning the joint distribution of a complex distribution over an expanded variable space is a difficult task. Latent variable models have been the recommended and well-established schemes in literature for this problem. These models are also used for dimensionality reduction and better representation of data.
- Visualization of unlabeled multidimensional data is another unsupervised task. In this research, we have explored the dimensionality reduction as an underlying scheme for developing better multidimensional data visualization tools.

Applications of Unsupervised Learning in Networking

The main applications of unsupervised learning are:

- Clustering
- Visualization
- Dimensionality Reduction
- Finding Association Rules
- Anomaly Detection

Clustering is the process of grouping the given data into different clusters or groups. Unsupervised learning can be used to do clustering when we don’t know exactly the information about the clusters. Clustering automatically split the dataset into groups base on their similarities.

Visualization is the process of creating diagrams, images, graphs, charts, etc., to communicate some information. This method can be applied using unsupervised machine learning.

Dimensionality reduction is the process of reducing the number of random variables under consideration by getting a set of principal variables. One way to do dimensionality reduction is to merge all those correlated features into one. This method is also called feature extraction.

In association rule learning, the algorithm will deep dive into large amounts of data and find some interesting relationships between attributes. This is the process of finding associations between different parameters in the available data. It discovers the probability of the co-occurrence of items in a collection, such as people that buy X also tend to buy Y.

Anomaly detection is the identification of rare items, events or observations which brings suspicions by differing significantly from the normal data. In this case, the system is trained with a lot of normal instances. So, when it sees an unusual instance, it can detect whether it is an anomaly or not.

Disadvantages of Unsupervised Learning in Networking

- Precise information regarding data sorting, and the output as data used in unsupervised learning is categorized and not identified.
- A reduced amount of accuracy of the results is because the input data is not known and not labeled by people in advance. This means that the machine requires to do this itself.
- The spectral classes do not always correspond to informational classes.
- The user needs to spend time interpreting and label the classes which follow that classification.

- Spectral properties of classes can also change over time so you can't have the same class information while moving from one image to another.

Conclusion

In this paper, a comprehensive survey of machine learning tasks, latest unsupervised learning techniques, and trends, along with a detailed discussion of the applications of these techniques in networking related tasks are been discussed.

Due to the versatility and evolving nature of computer networks, it was impossible to cover each and every application; however, an attempt has been made to cover all the major networking applications of unsupervised learning and the relevant techniques.

Despite the recent wave of success of unsupervised learning, there is a scarcity of unsupervised learning literature for computer networking applications, which this paper review aims to address.

References

1. Usama, Muhammad, et al. " Unsupervised machine learning for networking: Techniques, applications and research challenges." IEEE Access 7 (2019): 65579-65615.
2. G. Casolla, S. Cuomo, V. S. d. Cola and F. Piccialli, " Exploring Unsupervised Learning Techniques for the Internet of Things," in IEEE Transactions on Industrial Informatics, vol. 16, no. 4, pp. 2621-2628, April 2020.

3. O. Kotlyar, M. Pankratova, M. Kamalian, A. Vasylchenkova, J. E. Prilepsky and S. K. Turitsyn, " Unsupervised and supervised machine learning for performance improvement of NFT optical transmission," 2018 IEEE British and Irish Conference on Optics and Photonics (BICOP), London, United Kingdom, 2018, pp. 1-4.
4. https://pythonistaplanet. com/applications-of-unsupervised-learning/
5. https://machinelearningmastery. com/ super vised-and-unsupervised-machine-learning-algorithms/
6. https://www. frontiersin. org/articles/10. 3389/ fncom. 2019. 00031/full
7. https://www. guru99. com/unsupervised-mach ine-learning. html