A Review of Classification Techniques in Machine Learning

Jyoti Pipariya
Department of Electronics & Communication Engineering
Acropolis Institute of Technology and Research, Indore, Madhya Pradesh – 453771, India

Abstract— The data set used by machine learning algorithms, every instance, is represented using the same set of features. These features may be binary, continuous, and categorical. On the basis of features, it may be classified as supervised and unsupervised learning. If instances are given with known labels which correspond to the correct outputs then the learning is said to be supervised. Whereas unsupervised learning corresponds to the instances which are unlabeled. A special kind of learning called reinforced learning is in which the training information that is provided to the learning system by the external environment is in the form of scalar reinforcement signal. This signal constitutes a measure on the system operation.

Keywords: Machine Learning, Data Analysis, Machine Learning Algorithms, Supervised Machine Learning, Support Vector Machine

I. Introduction

Past few years have witnessed an unmatched rise of interest in machine learning due to the virtue of large datasets available. In addition, remarkable advancement in algorithms and exponential growth in computing skills further contributed in drawing attention towards machine leaning. Today, machine learning algorithms are brilliantly utilized for an array of tasks such as classification, clustering, regression, and dimensionality reduction of huge data sets [1]. Machine learning, in fact, has proved to outperform humans in various disciplines. Thus, several spheres of our day to day life are driven by machine learning, such as web-searches [2], image and speech recognition [3, 4], fraud detection [5], email or spam filtering [6], credit score [7] and so on.

Machine learning has already been applied in the field of biology [8] and chemistry [9] from fairly long time, and recently its importance has also increased in the field of solid-state materials science [10]. Conventionally, experiments were the key in discovering and characterizing new materials. Experimental research was conducted over a long time period for an extremely limited number of materials, as it enforces high prerequisites in terms of resources and equipment. Due to these limitations, important discoveries were made mostly through human intuition or even coincidence [11]. While the new wave of promises and innovations around machine learning falls short of the requirements that propelled early data-driven artificial intelligence research [12, 13], learning algorithms have proven to be useful in a number of important applications.

This article provides a very brief introduction to key concepts in machine learning and the recent advances. We aim at highlighting conditions under which the use of machine learning is justified in engineering problems, as well as specific classes of learning algorithms that are suitable for their solution.

II. Categories of Classification Algorithms

Machine learning is a subcategory of artificial intelligence that provides computers the ability to automatically learn and constantly improve from the experience without the need of any unambiguous programming. By means of machine learning, computers replicate human learning activities, develop self-improvement methods, recognize existing knowledge, and acquire new knowledge and new skills in order to continuously improve the performance and achievement. It basically enables the computers to make data-driven decisions and algorithms are designed in such a way that there is constant learning and improvement process over the time when exposed to new data. This makes machine learning much faster than human learning and also the spread of acquired knowledge is easier. Therefore, the progress made by humans in the field of machine learning boosts the proficiency of computers which in turn greatly impacts the human society.

III. Basic model of machine learning

A very basic model can easily explain process of machine learning [14]. In this whole process, the prime factor is the quality of information provided by the external environment (Fig. 1). Thus, the external environment represents the source of outside information set that is served in a particular form. This information set is then processed into knowledge during the learning process and the knowledge is kept in a repository. Repository has collection of a number of general principles that act as guidelines for the implementation process. Since the environment provides all sorts of information for the learning process, the quality of information has a direct impact on the learning comprehension and determines whether the process will be easy or challenging.

The second factor that has a direct impact on learning system is repository. There are different types of repositories that represent various ways of expression of knowledge and each of these has its own forte. Few examples of such fashion of expressions are eigenvector, logic statements of the first order, production rules, semantic networks and frameworks etc.

The process of learning begins with observation of data with the aim of looking for patterns in data and make better decisions in the future based on the models that we have provided. The main aim is to allow the computers learn automatically without human involvement and adjust actions appropriately. When new input data is introduced to the machine learning algorithm, it makes a prediction on the basis of the model. The obtained prediction is then analyzed for accuracy and if the accuracy is within acceptable limits, the machine learning algorithm is implemented. In case if the accuracy is not acceptable, the machine learning algorithm is trained again and again with an extended training data set.

The next step in the process is implementation. It is a phase in which the knowledge stored in repository is used to finish an assigned task. This is followed by providing the feedback of information gathered during task completion to the learning process step. This facilitates the guidance of whole process. The four aspects that should be taken into consideration for completion of the process are potency of expression, ease in deriving inference, ease in repository modification, and the ease in escalating the derived knowledge.

The process of applying supervised machine learning to a real world problem involves the collection of data set, identification of the required data, data pre-processing, defining a training set, algorithm selection and evaluation of test set. For the collection data set, if a requisite expert is available then the attributes and features which are most informative may be collected. This is similar to the fetching of the information needed from the data to make a good judgment for choosing a class model.

The second step is data pre-processing which features into format that is suitable for the estimators. Generally, machine learning model prefer standardization of the data set. Depending on the circumstances, researchers have a number of methods to choose from to handle missing data [15]. Instance selection is not only used to handle noise but to cope with the infeasibility of learning from very larger data sets. Instance selection in these data sets is an optimization problem that attempts to maintain the mining quality while minimizing the sample size [16]. Feature subset selection is the process of identifying and removing as many irrelevant and redundant features as possible [17].

IV. Algorithm Selection

In general, classification algorithms are classified as follows:

- Linear Classifiers- This includes logistic regression, naïve bayes and Fisher’s linear discriminant.
- Support Vector Machines- This includes least square support vector machines.
- Kernel Estimation- This includes k-nearest neighbor.
- Quadratic Classifiers
- Decision Trees- This includes random forests.
- Neural Networks
- Learning Vector Quantization

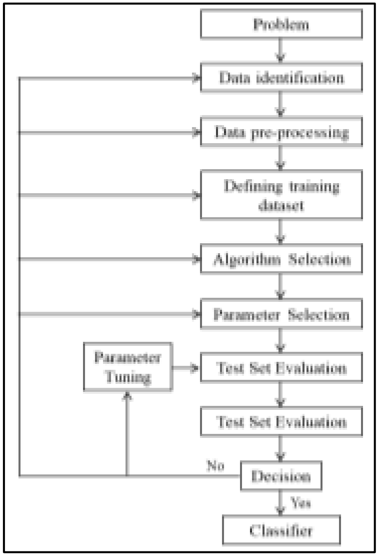

Fig. 1: Basic model of machine learning

V. Machine Learning Methods

Machine learning algorithms are often classified as supervised or unsupervised.

- Supervised machine learning algorithms: For supervised algorithms, we are required to provide both input and desired output, followed by supplying feedback about the accuracy of predictions during algorithm training. The learning algorithm can also compare its output with the accurate, envisioned output and find possible errors in order to modify the model accordingly. After the completion of initial learning process, the variables or features and the models used are analyzed and the most potent one is finalized to develop predictions. Once the training is complete, the algorithm will apply the same learned features and models to new input data.
- Unsupervised machine learning algorithms: There is no need to train unsupervised algorithms with desired outcome data. During the process, the model learns through observation and finds hidden structures in the data. Once the model is given an input dataset, it automatically finds patterns and relationships in the input dataset by creating clusters in it. However, it cannot add labels to the clusters. Thus, unsupervised algorithms are used when the information used to train is neither classified nor labeled. In fact, an iterative approach called deep learning is used to review data and arrive at conclusions. Once trained, the algorithm can use its bank of associations to interpret new data. Unsupervised learning algorithms, also called neural networks, are used for more complex processing tasks than supervised learning systems. These tasks include image recognition, speech-to-text and natural language generation. But these algorithms are competent only in the age of big data, as massive amounts of training data is a mandate.
- Semi-supervised machine learning algorithms: It lies in between supervised and unsupervised learning. Semi-supervised algorithms employ both labeled and unlabeled data for training. Usually, a small amount of labeled data is taken along with a large amount of unlabeled data. With the help of this method, the systems are able to considerably improve learning accuracy. In general, semi-supervised learning is employed when the acquired labeled input data requires competent and relevant resources for training. On the other hand, unlabeled input data usually doesn’t require additional resources.
- Reinforcement machine learning algorithms: This method involves interaction with the environment in order to figure out the best outcome. It basically adheres to the concept of hit and trial method. The most relevant characteristics of reinforcement learning are trial, error search and delayed reward. A simple reward feedback is required for the agent to learn which action is the best. Thus, the agent is rewarded or fined with a point for each correct or a wrong answer, respectively. Now, on the basis of all the positive reward points obtained the model trains itself. And again once trained it gets ready to predict the new data within a specific context in order to maximize its performance.

A. Decision Tree Algorithm:

This algorithm splits the datasets into two or more homogenous sets based on the most significant attributes making the groups as distinct as possible [18]. These are the trees that classify the instances by sorting them based on feature values. Each node represents a feature in an instance to be classified, and each branch represents a value that the node can assume.

B. Logistic Regression:

It estimates discreet values (binary values like true/false) based on the given set of independent variables. The values obtained always are in the form of 0 and 1 since it predicts the probability.

C. Support Vector Machine (SVM):

SVM is based on the notion of a “margin”-either side of a hyperplane that separates two data classes. Maximizing the margin and thereby creating the largest possible distance between the separating hyperplane and the instances on either side of it has been proven to reduce the upperbound on the expected generalization error [18].

VI. Applications of Machine Learning

Today, machine learning is being used in a wide array of applications. The most distinguished example is News Feed on various platforms, which uses machine learning to personalize each member's feed. The software simply practices statistical analysis and predictive analytics to identify patterns in the user's data and use those patterns to cluster the News Feed. Depending on the user’s activity the new data will be included in the data set and News Feed will be fine-tuned consequently.

Machine learning is also participating in a range of enterprise applications. Customer relationship management (CRM) systems use learning models to analyze email and persuade sales team members to respond to the most important messages first. More advanced systems can even recommend potentially effective responses. Business intelligence (BI) and analytics vendors use machine learning in their software to help users automatically identify potentially important data points. Human resource (HR) systems use learning models to identify characteristics of effective employees and rely on this knowledge to find the best applicants for open positions.

Machine learning also plays an important role in self-driving cars. Deep learning neural networks are used to identify objects and determine optimal actions for safely steering a vehicle down the road.

Virtual assistant technology is also powered through machine learning. Smart assistants combine several deep learning models to interpret natural speech, bring in relevant context -- like a user's personal schedule or previously defined preferences -- and take an action, like booking a flight or pulling up driving directions.

VII. Conclusion

Just like the industrial revolution, in the field of machine learning machines are progressively trained to identify patterns and to find relations between properties and features more efficiently than us. Machine learning enables analysis of big data. While it generally delivers faster, more accurate results in order to identify profitable opportunities or dangerous risks, it may also require additional time and resources to train it properly. Combining machine learning with artificial intelligence and cognitive technologies can make it even more effective in processing large volumes of information.

References

[1] Marsland S. Machine Learning (CRC Press, Taylor & Francis Inc., Boca Raton, FL, 2014.

[2] Pazzani M, Billsus D. Learning and revising user profiles: the identification of interesting web sites. Mach. Learn, 1997; 27: 313–331.

[3] Liu SS, Tian YT. Facial expression recognition method based on gabor wavelet features and fractional power polynomial kernel PCA. In Advances in Neural Networks - ISNN 2010 (eds Zhang, L., Lu, B.-L. & Kwok, J.) 144–151 (Springer, Berlin, Heidelberg).

[4] Waibel A, Lee KF. (eds) Readings in Speech Recognition (Morgan Kaufmann, Burlington, MA, 1990).

[5] Chan PK, Stolfo SJ. Toward scalable learning with non-uniform class and cost distributions: a case study in credit card fraud detection. In KDD’98 Proc. Fourth International Conference on Knowledge Discovery and Data Mining (eds Agrawal, R., Stolorz, P. & Piatetsky, G.) 164–168 (AAAI Press, New York, NY, 1998).

[6] Guzella TS, Caminhas WM. A review of machine learning approaches to spam filtering. Expert Syst. Appl. 2009; 36: 10206–10222.

[7] Huang CL, Chen MC, Wang CJ. Credit scoring with a data mining approach based on support vector machines. Expert Syst. Appl. 2007;33: 847–856.

[8] Baldi P, Brunak S. Bioinformatics: The Machine Learning Approach (The MIT Press, Cambridge, MA, 2001).

[9] Noordik JH. Cheminformatics Developments: History, Reviews and Current Research (IOS Press, Amsterdam, 2004).

[10] Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 2019; 5: 1-36.

[11] Rajan K. Materials informatics. Mater. Today, 2005; 8: 38–45.

[12] Levesque HJ. Common Sense, the Turing Test, and the Quest for Real AI: Reflections on Natural and Artificial Intelligence. MIT Press, 2017.

[13] Pearl J, Mackenzie D. The Book of Why: The New Science of Cause and Effect. Basic Books, 2018.

[14] Zeng-bo AN, ZHANG Yan. The Application Study of Machine Learning [J]. Journal of Changzhi University, 2007; 24: (2):21-24.

[15] Batista G, Monard MC. An analysis of four missing data treatment methods for supervised learning, applied artificial intelligence, 2003; 17: 519-533.

[16] Liu H, Motoda H. Instance selection and constructive data mining, Kluwer, Boston 2001.

[17] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. JMLR, 2004; 5:1205-1224.

[18] Kotsiantis SB. Supervised machine learning: A review of classification techniques. Informatica. 2007; 31: 249-268.