THINK INDIA JOURNAL
ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

A Comparative Study and Analysis of Classification Methods In Machine Learning

Goutam Majumder[1], Richa Jain[2]

School of Computer Science & Engineering

Lovely Professional University, Jalandhar, India – 144411

[1]goutam.23320@lpu.co.in, [2]richa.17688@lpu.co.in

Abstract – In this paper, we have highlighted the importance of machine learning algorithms in our daily routines. In addition, with their learning mechanisms, we havereported various classification algorithms. Thereafter various machine learning models are documented with algorithms used by them. Besides this, a general overview of machine learning and its various open source tools and machine learning applications are also reported. A comparative analysis with other related areas such artificial intelligence and deep learning also reported.

keywords – Machine Learning, Classification, Regression, Scikit Learn, Keras.

1. INTRODUCTION

Machine Learning (ML) is an artificial intelligence sub-set that uses computer algorithms to learn from data and information autonomously. Based on neuro-psychological learning, ML was introduced by Hebb in 1949 as Hebbian Learning theory. At present ML is something, which is usedto solve the problems related to our daily life. Machine learning is used in Virtual Personal Assistants (e.g. siri, Alexa), which helps us to provide any assistance. In road traffic predictions and online transportation network, ML helps us to make predictions while commuting. In addition to these, it has various other applications in video Surveillance, social media services(facial recognition, people you may know), email spam and malware filtering.

Certain tasks are extremely hard to configure by hand, like spam filtering, facial recognition, speech recognition, machine translation etc. Virtually all learning concerns can be described as mappings between input and output. Classification systems play an important role in decision-making activities by classifying the information available on the basis of certain parameters. We present numerous classification types used in machine learning, such as binary, multiclass and multi-label. Depending on the type of problem and the type of data used for input and output, machine learning types differs in many ways.Therefore, learning algorithms are commonly categorized as supervised learning (learning from examples), unsupervised learning (discovering underlying patterns), semi-supervised learning (with marked and unlabeled data), enhancing learning (learning behavior by interacting in an environment). Specific data and issue-based models includes ANN, SVM, decision tree, Regression Analysis, Genetic Algorithms, and Bayesian Network.

Organization of the paper is as follows: Section 2 highlights various classification strategies, required learning mechanism are listed in Section 3. Section 4, reported various types of models used by machine learning algorithms. The impact of machine learning to daily life is

reported in Section 5. Section 6 provides the information related to the open source tools and softwares used by the researchers to build their machine learning model. At last conclusion of the paper is drawn in Section 7.

2. TYPES OF CLASSIFICATION

In machine learning a classification problem is categorised based on the number of class labels required to correctly identify for a problem instance.

1.1. Binary Classification

When predicted classes are label with maximum two classes, then classifications are called binary or binomial. Various methods are proposed to perform binary classification in statistical classification such as: decision tress, random forests, Bayesian networks, support vector machines, logistic regression. These methods are domain dependent like, random forest algorithm performs better than support vector machine for 3D data points[1].

1.2. Multiclass Classification

In machine learning, identifying a problem instance into one of three or more classes are called multi-class or multinomial classification. The multiclass classification algorithms categories into any of three techniques:

- Transforming to binary classification: two strategies are used to transform a multiclass classification problem to multiple binary class classification such as: one-vs-rest (OvR) and one-vs-one (OvO). The *OvR*strategy,uses a single classifier per class considered as positive class and others are considered as negative class[2].
  In *OvO* strategy, $K(K-1)/2$ numbers of binary classifiers are trained in K different ways and a voting scheme is applied to get the final prediction[2].
- Neural network based multiclass strategies: in the output layer instead of one binary classifier, multiclass perceptrons also used to predict a class label.

To address multiclass problem several algorithms are developed using support vector machine[3], Naïve Bayes and k – nearest neighbors.

1.3. Multi-label classification

Multi-label classification is a subset of multiclass classification problem, in which a problem instances are labelled with multiple types. In literature, a multi-label problem is transformed into one of these following variants:

- Transforming to a binary classification problem: in this category one independent classifier is used to train each label[4], [5]. For any unseen problem, it predicts the class label as combined output. For this purpose, a chain of classifiers such as Bayesian network is applied over a problem [6].

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

- Transforming to a multi-class classification problem: for this each combinations of class labels are represents by a classifier. Suppose A, B and C are three classes than label power set combinations are used to train a pattern [7].
- Ensembles Methods:each Label Powerset (LP) classifier is trained with a random subset of labels. Finally, voting schemes are used to produce a combined output[8].

3. TYPES OF LEARNING

Types of machine learning differs in two ways. At first it depends on problem type and secondly it depends on the data is used for input and output. In literature these learning algorithms are broadly categorise into following groups:

3.1. Supervised learning

It learnsa model on a set of data, which have the inputs with desired outputs [9]. This data is called training data and each data points can have one or more inputs and a desired output. A feature vector is formed by combining all the inputs and represented as *n*-dimensional array as matrix. Supervised learning algorithms learned optimized function, which will predict the desired data for an unseen data points are called test data[10],[5].
Supervised algorithms are used to solve problems like classification and regression. In classification predicted output will be from a desired set of output and it can be a binary or multiple classification problem. In regression it predicts an output as continuous values from a range such as predict the price of a house[2].

3.2. Unsupervised learning

Unsupervised learning methods learns patterns of data by forming groups and clusters among the inputs. The training data points only contains inputs no desired output is tag with data points. During testingthe model tries to map a new data points into one of clusters. In statistics, density estimations is one of the central applications of unsupervised learning[11].

3.3. Semi-supervised learning

It falls between supervised and unsupervised learning. The training data contains both labelled and unlabelled data in a combination of minimum to maximum of labelled and unlabelled data respectively. At first learning algorithm first forms the clusters of similar data and then data labels are used to label the unlabelled data. The typical applicationsare speech analysis, content classification and protein sequence classification[12].

3.4. Reinforcement learning

In this type of learning software agents take actions in an environment to maximize the output. This technique differs from supervised learning, in which labelled input and output needs to be provided. It focuses in to maintaining a balance between exploration and exploitation of the knowledge gathered from environment[13].

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

Markov decision process is used to formulate an environment and in literature many reinforcement learning algorithms uses dynamic programming for this purpose. Due to this principle, reinforcement learning techniques also popular among other disciplines such as multi-agent systems, genetic algorithm, operation research and game theory.

4. MODELS USED IN MACHINE LEARNING

In machine learning efforts are put to build intelligent models, which is trained on data and then predict the output on unseen data. Researchers are developed and used various types of models based on the data and problem they want to address.

4.1. Artificial Neural Network

It is also called cennectionist network inspired from biological neural network. These system are learn to perform a job, by cosidering example, without programming with job specific rules. It has varous component and these components are together to provide human like performance. These compnents are as follows:

- Neurons: it receives an input signal and combined it with internal state and threshold using an activation function. At last it produce an output using an output function. The input signals can be an image or sequences of characters and final output can be a categorical value or a random value.

- Connections and weights: neurons of neural networks have connections and output of one neuron is transmitted to another neuron as input signal. A neuron can have multiple input and output.

- Propagation functions: these functions computes the inputs for one neuron as weighted sum of all outputs of the previous layer neurons.To minimize the error between the actual output and the expected output, a bias term can be applied to the previuos output.

4.2. Decision Tree

It is commonly used in data mining and the goal is to predict the value for a target variable on several input variables. It is tree like representations of a classification problem, in which each leaf represents a value of a target variable.

4.3. Support Vector Machines

These are also referred to as support vector networks, a set of supervised learning algorithms used to classify and regress.These set of supervised learning algorithms consits of binary, linear classifier and non-probabilistic methods and methods like Platt scaling also used by SVM as probabilistic classifiction techniques. SMVs are also used for non-linear classification and various kernel tricks are used by mapping the inputs into higher-dimensional feature spaces[14].

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

Kernel methods are considered as instance-based learners, in which SVMs remember the i[th] input (x_i, y_i) and learn for a corresponding weight w_i.

4.4. Regression Analysis

As a regression analysis, a wide variety of statistical methods are developed that finds the relationship between input values and their associated characteristics. Linear regression is the most common form of regression analysis, where a line is drawn to best fit the data according to mathematical criteria such as the least squerd error, root mean squerd error.

Logical regression is another variety of regression approch which used for classification, which helps us to finds the non-linearty between the data uisng various kernel tricks. Regression model like polynomial regression is used by Microsoft Excel for trendline fitting[1].

4.5. Bayesian Network

These are the probabilistic graphical models trained from data or expert opinion. This model represnts data with their conditional dependencies and represents their relation using a direct acyclic graph (DAG).

Bayesian networks learns sequence of data such as text, protein sequences are called dynamic Bayesian network. If X is a Bayesian network of G, where G is a direct acyclic graph with V vertices and E edges, then its probability density functions can be defined as follows:

$$p(x) = \prod_{v \in V} p\left(x_v | x_{pa(v)}\right) \tag{1}$$

where $pa(v)$ is the set of parents of v (i.e. those vertices pointing directly to v via a single edge). For any set of randome values, chain rules are used to calculate the joint probability distribution as follows:

$$P(X_1 = x_1, \cdots, X_n = x_n) = \prod_{v=1}^{n} P(X_v = x_v | X_{v+1} = x_{v+1}, \cdots, X_n = x_n) \tag{2}$$

4.6. Genetic Algorithms

To identify the feature values from a set of inputs genetic algorithms are used. Techniques such as mutation and crossover are used to generate the new genotypes in the process of finding good solutions for a problem. Machine learning techniques are used to improve the performance of genetic and evolutionary algorithm [15].

[1] https://facultystaff.richmond.edu/~cstevens/301/Excel4.html

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

5. APPLICATIONS OF MACHINE LEARNING

Machine Learning is applied successfully over the several fields starting from agriculture, software engineering, online advertising, time series forecasting and in health care sector. In our day to day activities we take helps from machine learning models without knowing about this.

5.1. Virtual Personal Assistants

Various interactive applications allows users to enter text or record speech with an intelligent computational system. Information systems like, interactive voice response (IVR) interacts with their customers to provide a series of choices (by a recorded voice) [16]. Machine learning is important part of these intelligent systems, which improve their performance on the basis of previous involvement with their consumers[2].

5.2. Social Media Services

Social media platforms like Facebook, Twitters are used for product advertising and marketing purpose. These platforms are used to deliver the product promotions to the consumers. These type of product promotions.[3] Social Media Marketing (SMM) helps the business to reach a large number of consumers and gain brand recognition by posting product related ads to the consumer profile page directly.

5.3. Email spam and malware filtering

According to the report published by TopTenReviews, 40 percent of the emails considered as spam, which is sum of 12.4 billion email out of 31 billion email per day[4]. This site also listed best five spam filters available by 2019 and SpamBully[5] email filter rated highest as 10 as compared to other filters.

5.4. Product Recommendations

E-commerce companies like Flipkart and Amazon incorporates product recommendation engine with the profile page of their customers. These machine learning models uses the keywords available in the message sent by a participant in a discussion session. Further, product recommendation system can be categorise into any of these following types:

- Content based recommendation system: it learns from a user profile interest in certain features related to the product. It has limited used, because it is directly or indirectly related with the user profile. Before recommendations machine learning models need to be collect enough ratings from the user profile.

[2] https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0
[3] https://techterms.com/definition/smm
[4] https://www.toptenreviews.com/best-spam-filter
[5] https://spambully.com/

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

- Collaborative recommendation system: in this category rather finding the similarity between the user profile or the product a consumer purchased earlier with the newly available products, it recommends between the consumer who likes similar products[17].

6. TOOLS AND SOFTWARE'S AVAILABLE FOR MACHINE LEARNING

6.1. Scikit Learn

It is an open source machine learning library also known as sklearn for python programming language[6]. It has various classification and regression algorithms such as random forest, k-means clustering. It also supports various scientific libraries such as Numpy and Scipy.

6.2. PyTorch

It is an open source machine learning software based on a torch library for computer vision and applications for natural language processing[18]. It is developed by Facebook's AI Research Lab and released under the Modified BSD license as open source software. It supports two high-level functions such as computing tensors and deep neural networks[7].

6.3. TensorFlow

It is free and open source library used for dataflow and differential programming for various task[8]. It is found by Google Brain team for internal google use and released under Apache License 2.0. The machine learning community uses this to build deep artificial neural network models[18].

6.4. Weka

Waikato Environment for Knowledge Analysis (Weka[9]) founded by the University of Waikato, New Zealand. It is released under GNU General Public License with an associate book entitled "Data Mining: Practical Machine Learning Tools and Techniques".

6.5. KNIME

[6] https://scikit-learn.org/stable/
[7] https://pytorch.org/
[8] https://www.tensorflow.org/
[9] https://www.cs.waikato.ac.nz/ml/weka/

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

It is a data analytics, reporting and integration tool that is free and open source.It is mainly used in pharmaceutical research[19], customer data analysis and business analysis purpose[10].

6.6. Colab

Google Colab is a platform, which provides the platform to build a machine learning models. It support GPUs as well as TPUs and build on top of Jupyter free collaborative environment[11].

6.7. Keras.io

It is an open source neural network python library and capable of running with on top of various machine learning library such as TensorFlow, R and Theano etc. It is part of a research project ONEIROS and François Chollet is a Google engineer is the primary author of it[12].

6.8. Rapid Miner

It is an open source data science library formally known as YALE (Yet Another Learning environment) was developed by R.Klinkenberg et al., in early 2001. It is released use AGPL license with non-open source components. It provides an integrated environment to perform deep learning, data analytics activities[20].

7. CONCLUSION

In this paper, we have commented on various machine learning strategies in this paper. We also specifically highlighted the different types of classification system that artificial intelligence community uses. We also mentioned numerous open source libraries available for building and creating models for machine learning. It also discusses the effects of machine learning in our daily lives.

References

[1] Y. Lu and C. Rasmussen, "Simplified markov random fields for efficient semantic labeling of 3d point clouds.," in *International Conference on Intelligent Robots and Systems*, 2012.

[2] C. M. Bishope, Pattern Recognition and Machine Learnin, Springer, 2006.

[3] H. Yu, J. Han and K. C.-C. Chang, "PEBL: positive example based learning for web page classification using SVM.," in *Proceedings of the eighth ACM SIGKDD*

[10] https://www.knime.com/
[11] https://colab.research.google.com/
[12] https://keras.io/

international conference on Knowledge discovery and data mining, 2002.

[4] J. Read, B. Pfahringer, G. Holmes and E. Frank, "Classifier chains for multi-label classification," *Machine learning*, p. 333, 2011.

[5] J. Read, L. Martino and D. Luengo, "Efficient monte carlo methods for multi-dimensional learning with classifier chains," *Pattern Recognition*, vol. 47, no. 3, pp. 1535-1546, 2014.

[6] O. Soufan, W. Ba-Alawi, M. Afeef, M. Essack, P. Kalnis and V. B. Bajic, "DRABAL: novel method to mine large high-throughput screening assays using Bayesian active learning," *Journal of cheminformatics*, vol. 8, no. 1, p. 64, 2016.

[7] N. SpolaôR, E. A. Cherman, M. C. Monard and H. D. Lee, "A Comparison of Multi-label Feature Selection Methods using the Problem Transformation Approach," *Electronic Notes in Theoretical Computer Science*, vol. 292, pp. 135-151, 2013.

[8] T. Grigorios and I. Vlahavas, "Random k-Labelsets: An Ensemble Method for Multilabel Classification," in *In European conference on machine learning*, 2007.

[9] S. J. Russell and P. Norvig, Artificial intelligence: a modern approach, Malaysia: Pearson Education Limited, 2016.

[10] M. Mohri, A. Rostamizadeh and A. Tal, Foundations of Machine Learning, MIT Press, 2018.

[11] Neural Networks, In Allen B. Tucker (ed.) Computer Science Handbook, Second Edition (Section VII: Intelligent Systems) ed., Boca Raton, Florida: Chapman & Hall/CRC Press LLC, 2004.

[12] O. Chapelle, B. Schölkopf and A. Zien, Semi-supervised learning, Cambridge: MIT Press, 2006.

[13] L. P. Kaelbling, M. L. Littman and A. W. Moore, "Reinforcement learning: A survey," vol. 4, pp. 237-285, 1996.

[14] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[15] Z.-h. Zhan, Y. Lin , Y. Li, Y.-j. Gong, N. Chen and J.-h. Zhong, "Evolutionary Computation Meets Machine Learning: A Survey," *Computational Intelligence Magazine*, vol. 6, no. 4, pp. 68-75, 2011.

[16] O. Yadgar, N. Yorke-Smith, B. Peintner, G. Tur, N. Ayan, M. J. Wolverton, G. Acharya, V. S. Parimi, W. S. Mark, W. Wang and A. Kathol, "Generic virtual personal assistant platform". US Patent US 9082402 B2, 14 July 2015.

ISSN: 0971-1260
Vol-22-NO-30 Special Issue- December-2019

[17] F. O. Isinkaye, Y. O. Folajimi and B. A. Ojokoh, "Recommendation systems: Principles, methods and evaluation," *Egyptian Informatics Journal* , vol. 16, no. 3, pp. 261-273, 2015.

[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin and S. Ghemawat, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2016.

[19] A. Tiwari and A. K. Sekha, "Workflow based framework for life science informatics," *Computational Biology and Chemistry,* vol. 31, no. 5-6, pp. 305-319, 2007.

[20] M. Hofmann and R. Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications, CRC Press, 2013.

[21] S. Yegulalp, "Facebook brings GPU-powered machine learning to Python," *InfoWorld,* 19 January 2017.