TEST
Engineering & Management

A Research on Recent Machine Learning Access and Representation

Dhanalakshmi S, Indhumathi S

Dhanalakshmi S, Assistant Professor, Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore,Tamil Nadu.(E-mail: dhanalakshmis@skasc.ac.in,)
Indhumathi S,Assistant Professor, Department of Software Systems, Sri Krishna Arts and Science College, Coimbatore,Tamil Nadu.(E-mailindhumathis@skasc.ac.in)

Article Info
Volume 81
Page Number: 4589 - 4593
Publication Issue:
November-December 2019

Article History
Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *23 December 2019*

Abstract:

The technology advancement in computer field has made abundance of data to be stored, maintained and retrieved easily. The data stored can be in structured and unstructured form. These data can be used to predict useful knowledge. Lots of research work is carried out in machine learning which is derived as a branch of artificial intelligence. The machine learning algorithms are developed to allow system to evolve and predict the knowledge based on the given empirical data. The self-learning algorithms are developed to automate knowledge prediction. The classical machine learning algorithms can be classified as supervised, unsupervised, and semi-supervised learning. Machine learning algorithms are applied in various applications such as object detection, object recognition, face detection, image segmentation, economical and commercial usage, etc. In this work carried a study on the machine learning approaches that are widely used.

Keywords—*Machine Learning, Ensemble, Active, Cost-Sensitive.*

I. INTRODUCTION

In traditional computing a set of explicitly coded programs are used for problems solving. Machine learning differs from traditional computing. Machine learning facilitates the computer system to train the sample data and to create a model based on this data. These models are used to predict automatically in the future, based on the empirical data. The efficient machine learning algorithms should predict useful knowledge from trained data.

There are three main components to be considered for all machine learning algorithms. They are representation, evaluation and optimization. The representation indicates how the algorithm will represent the knowledge. The evaluation component indicates how to evaluate the hypotheses and the third component is optimization.

II. CLASSICAL MACHINE LEARNING ALGORITHMS

Depending upon the nature of learning system the classical machine learning is categorized as supervised learning, unsupervised learning and semi supervised learning [1].

Supervised Learning

This type of algorithms has a pre-determined class label value (eg: yes/no) for the trained data. The supervised learning algorithms further try to predict the target output values for the test data based on the previously learned dataset. So this type of learning algorithms have predictive model with labeled data. Supervised learning problems can be further categorized as classification and regression. The supervised learning algorithms which focus more with classification includes decision tree, random forest, Naïve Bayes, support vector machine, logistic regression [9].

Unsupervised Learning

This type of algorithms acts on the datasets without any direction or guidance. Unsupervised machine learning tries to predict the output using the dataset that is not labeled. This algorithms group the data based on the patterns and similarities without any trained data. Unsupervised learning problems can be further categorized as Clustering and Association. Unsupervised clustering types falls as hierarchical clustering, k-means clustering, DBScan, optics, agglomerative, divisive.

Semi Supervised Learning

In this type of learning only some of the data are labeled and most of the data are left unlabeled in the training dataset. Combinations of supervised and unsupervised approaches are used in semi supervised learning. The methods for semi supervised learning includes self training, co-training, multi view training and graph based models [10].

III. MACHINE LEARNING APPROACHES& RESULTS

In the current scenario, the machine learning research focuses on tremendous learning methods apart from the classical machine learning algorithms.

In this paper addressed some of the widely used machine learning approaches in the research.

- Ensemble learning methods
- Active learning
- Reinforcement learning
- Cost-sensitive learning
- Collective classification

Ensemble Learning Methods

An ensemble learning method uses multiple learners to solve the problem. It is a combination of multiple models. Ensemble model constructs multiple models from the data samples and all the predictions from multiple models are aggregated to predict the final output [3]. The prediction can be affected by various factors like difference in population, hypothesis and modeling techniques. The ensemble machine learning method is predicted in Fig 1.

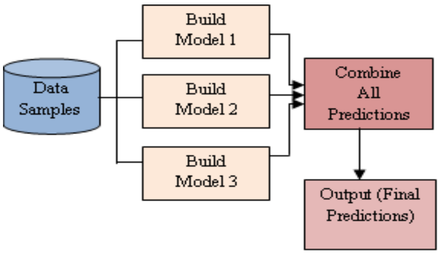


Fig 1.Ensemble Machnie Learning Method

For example if there is a need is to predict the quality of a product before purchasing, it can be learned by gathering multiple customers and experts opinion. Based on the multiple learners knowledge and their perspective can helps to retrieve useful information. These information can be aggregated to predict final decision about the product.

The possibilites of errors in ensemble learning can be categorized as Bias error and variance error.

Bias error is used to quantify how much the predicted value will be different from original value on an average.

Variance error quantifies the difference between each learning observation.

Commonly used techniques in ensemble learning are Bagging, Boosting and Stacking shown in Fig 2.

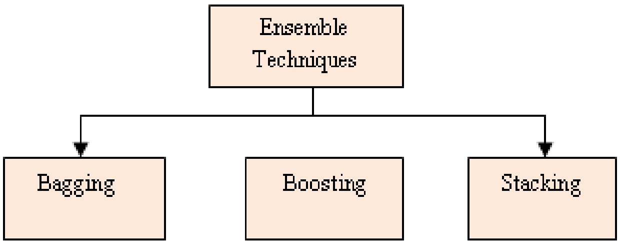


Fig 2. Commonly used Ensemble Learning Techniques

Bootstrap AGGregaING(Bagging) ensemble learning method choose N sample populations from the training data and train the classifier.The process is repeated until the desired size is achieved. The method calculates the mean of all predictions. The bagging approach helps to reduce the variance error.

Boosting is a repetition approach with weight adjustment based on previous classification. If observed classification is incorrect, the weight value is adjusted and observed again. Boosting helps in reducing bias error.

In stacking different classifiers are trained with different training set and different learners output is merged by taking average in the next level. This approach helps in reducing both variance and bias error depending upon the learner used.

Active Learning

Active learning falls under semi-supervised machine learning. The learning is performed by interactively querying the information source or user to retrieve the expected output. The advantage of active learning is the entire data is not needed to be labeled; it will be the user or learner task to request for relevant label [4].

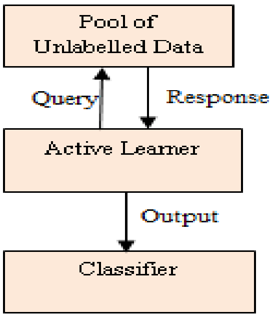


Fig 3. Active Learning

Fig 3. Shows that the active learning is performed by the following steps:

1. The learning starts with unlabeled data
2. With the help of expert pick few points at random and retrieve the labels
3. Fit the classifier to this labels
4. Repeat Step 2 and 3.

The two main components used in active learning are

- Oracle
- Query System

Oracle: This may be a human annotator or information acquisition system

Query System: The query system post queries to oracle for retrieving labels.

Different strategies are followed to post the queries to the learner. The main three active learning scenarios are shown in Fig 4.

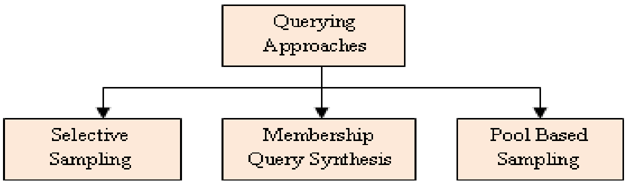


Fig 4. Active Learning Scenarios

Selective Sampling

In selective or streamline sampling the unlabeled samples are drawn one at a time from the data distribution. The learners must decide whether to query and also should decide whether to be labeled or to discard.

Membership Query Synthesis

In this scenario the learner request labels from instances in the entire space. This approach usability purely depends on the scenario because the oracle can have difficulties in interpreting label instances.

Pool Based Sampling

It assumes that there is an availability of pool instances to post queries to interpret the labels.

Reinforcement Learning

Reinforcement Learning learns by interacting with the environment. In reinforcement learning the decision is dependent on environment i.e. it is learned from experience. The output for the task is explored based on the state of current input and also use evaluation feedback called reward to redefine the behavior [2].

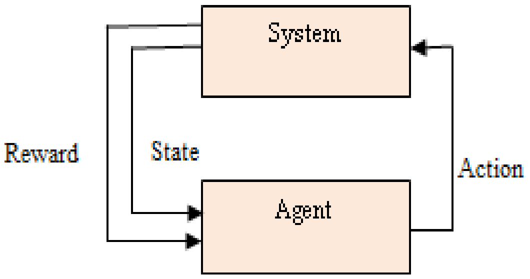


Fig 5: Reinforcement Learning

The above diagram Fig 5 predicts that this type of learning does not know what will be the input data and output data. It learns by exploring the dynamic environment by receiving rewards, states and uses the Action [5].

Reinforcement learning approaches in this context can be categorized as direct and indirect learning and is depicted in Fig 6.

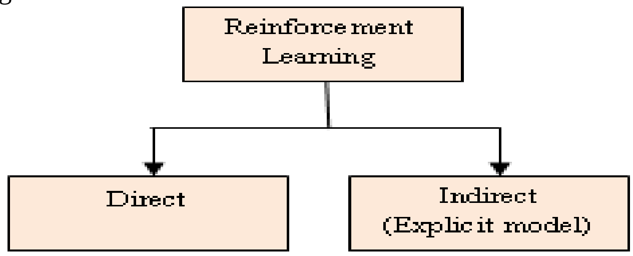


Fig 6. Learning Approaches-Classification

Indirect Learning

An explicit model of the environment is first estimated and an optimal policy for the estimated model is computed.

Direct Learning

The optimal policy is learned without an explicit model such schemes include search in policy space and value function based learning.

Cost-Sensitive Learning

Cost-Sensitive Learning takes the misclassification cost into consideration. The goal of learning with classification cost is to acquire a high accuracy during classification.

Classification is a important task in machine learning. Many classification algorithms has been developed and most of the algorithms is focused to minimize the error rate i.e. focus on minimizing the percentage of incorrect prediction of class labels. But all this algorithms does not consider the misclassification errors. But cost-sensitive learning takes misclassification cost into consideration.

Actual Class	Predicted Class		
		Positive	Negative
	Positive	0	C(FN)
	Negative	C(FP)	0

Table 1.Classification Cost Matrix

Total Cost (Misclassification) = FN x C(FN) + FP x C(FP)

FP – Number of wrongly predicted positive observations

FN – Number of wrongly predicted negative observations

C(FN) and C(FP) – Costs associated with FN and FP.

Cost matrix provides information about misclassification cost and it is shown in Table 1. The goal of this learning method is to choose the classifier with lowest total cost

Cost-sensitive learning can fall into two categories. First one is the direct method in which classifiers are designed with cost-sensitive.

The second category of learning is wrapper method also called as cost-sensitive meta learning method [6] and shown in Fig 7. The wrapper or meta learning method is also further classified as stratification, thresholding and sampling

Simple approaches for cost-sensitive learning are

- According to costs, the instances are re-sampled.
- According to costs, weighting of instances.

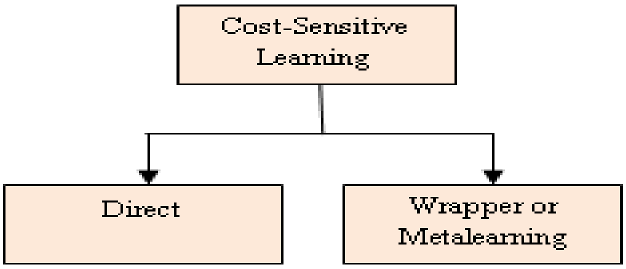


Fig 7: Cost-Sensitive Learning Categories

Stratification: In training data the frequency of classes are changed in proportion to their cost.

Sampling: In sampling, the distribution is distorted either by under-sampling or over-sampling. In under-sampling the data available for learning is reduced. In over-sampling, the learning time is increased.

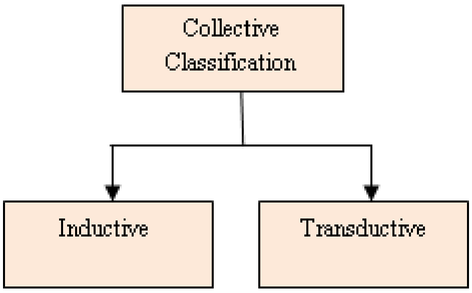
Collective Classification

Collective classification is a semi supervised learning method. Collective classification work is to assign correct label to the entire object in the network. Three types of correlations are used to assign the label of an object [7].

- Correlation between the observed attributes and the Object label.
- Correlation between the neighborhood observed attributes of object and the object label
- Correlation between the unobserved object labels in the neighborhood and the object label.

Thus collective classification relies on all the three types of correlation for combined classification to a set of interlinked objects .

Collective classification learning falls into two categories: Inductive and Transductive shown in Fig 8.



Inductive: In inductive learning, the label is assumed to be drawn from the distribution. The aim is to predict the newly drawn based on the labeled structures from the distribution.

Transductive: In Transductive learning, the domain in fixed with labeled set of data. The aim is to exactly predict the remaining instances.

The pros and cons of widely used machine learning methods are listed in Table 2.

Learning Approaches	**Pros**	**Cons**
Reinforcement Learning	Model is similar to the human being learning. Create a perfect Model. Can solve complex problems and achieve long-term results	Needs lot of data and computation Curse of dimensionality limits the model
Active Learning	Entire Data need not be labeled	Human annotator may face difficulties in interpreting instances in membership query synthesis
Ensemble Learning	Simple and Stable Accurate prediction results	High design and computation time
Collective Classification	Assign correct labels to the object	Over fitting issue
Cost Sensitive Learning	Misclassification cost is handled	Under-sampling results in higher variance

Table 2. Pros and Cons of Learning Approaches

IV. CONCLUSION

In this paper addressed the concepts of various machine learning approaches. Apart from classical machine learning algorithms like supervised and unsupervised learning, a lot of machine learning has been devised in the current research works. In this paper addressed such learning like ensemble learning, cost sensitive learning, active learning, collective classification and reinforcement learning. Each learning method is analyzed with the techniques adopted to implement the learning. The paper also highlights the pros and cons of widely used machine learning methods. This work gives the basic idea about the machine learning algorithms adopted in recent research works.

REFERENCES

1. Das, K., &Behera, R. N, "A survey on machine learning: concept, algorithms and applications", International Journal of Innovative Research in Computer and Communication Engineering, vol.5(2), pp.1301-1309, 2017.
2. Musumeci, F., Rottondi, C., Nag, A., Macaluso, I., Zibar, D., Ruffini, M., &Tornatore, M, "An overview on application of machine learning techniques in optical networks", IEEE Communications Surveys & Tutorials, vol.21(2), pp:1383-1408, 2018.
3. Dietterich, T. G, "Ensemble methods in machine learning", International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg,pp:1-15,2000.
4. Aggarwal,C,C.,Kong, X., Gu, Q., Han, J., & Philip, S.Y, "Active learning: A survey, Data Classification, Chapman and Hall (CR), pp: 599-634, 2014.
5. Sammut, C., & Webb, G. I. (Eds.), "Encyclopedia of machine learning", Springer Science & Business Media,2011.
6. Ling, C. X., & Sheng, V. S, "Cost-sensitive learning and the class imbalance problem", Encyclopedia of Machine Learning: Springer, 2011.

7. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., &Eliassi-Rad, T," Collective classification in network data", AI magazine, vol. 29(3), pp. 93-93,2008.
8. Aggarwal, C. C. (Ed.), Data classification: algorithms and applications. CRC press,2014.
9. Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., &Akinjobi, J," Supervised machine learning algorithms: classification and comparison", International Journal of Computer Trends and Technology (IJCTT),vol. 48(3), pp.128-138, 2017.
10. Prakash, V. J., &Nithya, D. L, "A survey on semi-supervised learning techniques", arXiv preprint arXiv, pp:1402.4645,2014