A survey on machine learning and outlier detection techniques

ZeeshanAhmad Lodhia1[†] and Akhtar Rasool2[††], and Gaurav Hajela3[†††]

Maulana Azad National Institute Of Technology, Bhopal, Madhya Pradesh, India

Summary
Machine learning is one of the most popular field in the computer science having different types of techniques such as supervised learning, unsupervised learning, reinforcement learning and the various techniques which are lying under them, so in order to understand these different machine learning techniques a survey on these machine learning techniques has been done and tried to explain these few techniques. The machine learning techniques try to understand the different data sets which are given to the machine. The data which comes inside can be divided into two types i.e. labelled data and the unlabeled data. These have to tackle both of the data. Those techniques have been looked upon as well. Then the concept of outlier comes into picture. Outlier Detection is one of the major issues in Data Mining; to find an outlier from a group of patterns is a famous problem in data mining. A pattern which is dissimilar from all the remaining patterns is an outlier in the dataset. Earlier outliers were known as noisy data, now it has become very difficult in different areas of research. Finding an outlier is useful in detecting the data which can't be predicted and that which can't be identified. A number of surveys, research and review articles cover outlier detection techniques in great details. The paper discusses and it tries to explain some of the techniques which can help us in identifying or detecting the observation which show such kind of abnormal behavior, and in technical terms called as outlier detection techniques.

Key words:
Machine learning, Supervised learning, Reinforcement Learning, Unsupervised learning, Outliers

1. Introduction

Machine learning can extract knowledge from large amounts of data, and then the extracted knowledge for prediction. Machine Learning is, an ability for machines to learn from a training data, here a machine is built up to use certain algorithms through which it can take its own decisions and provide the result to the user. It is considered the subfield of Artificial Intelligence. Today Machine Learning is used for complex data classification and decision making [1]. In simple terms, it is the development of algorithms that enables the system to learn, and to make necessary decisions. It has strong ties to mathematical optimization that delivers methods, theory and application domain to the field and, it is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Certain examples applications are Spam filtering [2], optical character

recognition (OCR) [3], Search Engines [4] and Computer Vision [5].

Machine learning is a field of research that formally focuses on the theory, performance, and properties of learning systems and algorithms. It is a field build on ideas from many different kinds of fields such as artificial intelligence(AI), information theory, optimization theory, cognitive science, statistics, optimal control, and many other disciplines of science, engineering, and mathematics [6–9]. Machine learning has been implemented in a wide range of applications, it has covered approximately every scientific domain, that has brought a huge impact on the science and society [10]. Algorithm of machine leaning been used on different varieties of problems, including recommendation engines, informatics and data mining, recognition systems, and autonomous control systems [11].

Generally, the field of machine learning is divided into three subdomains:

1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

Supervised learning requires training with data which is labelled and it has inputs and desired outputs, whereas unsupervised learning does not require labeled data for training and it needs inputs without any desired outputs. Reinforcement learning learns from feedback received through interactions from an external environment. On the basis of the above three learning paradigms, a lot of application services and mechanisms have been proposed for dealing with data tasks [13–15]. For example, in [15], Google applies machine learning algorithms to retrieve huge amount of data obtained from the Internet for Google's translator, Android's voice recognition, image search engine and Google's street view.

The next important technique we come across is detecting of outliers. Outlier Detection is one of the important issues in Data Mining, detecting outliers from a group of patterns is one of the famous problem in data mining. Outlier is that kind of pattern which is different from all the other patterns in the data set. Outlier detection is very familiar

Manuscript received May 5, 2017
Manuscript revised May 20, 2017

area of research in data mining. It is very important task in different application domains. Earlier outliers were considered as noisy data, but it has now become a difficulty, discovered in various domains of research. The finding of outlier is useful in detection of in certain areas like fraud detection of credit cards, discovering computer intrusion and criminal behaviors etc.

The later part of the paper contains explanation on the different machine learning techniques, and explanation on outliers and outlier detection algorithms.

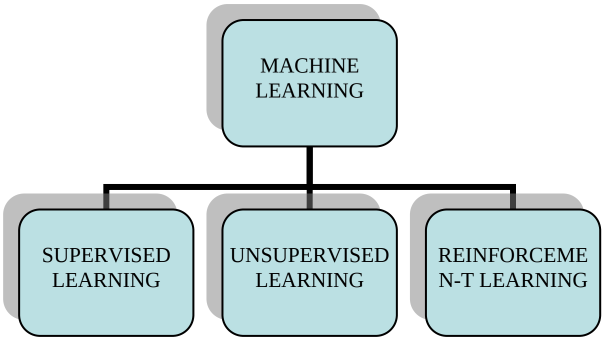

Fig. 1 Types of machine learning techniques

2. Supervised Learning

Supervised machine learning is the generation of algorithms which are able to produce general patterns and hypotheses by the help of externally supplied instances and predict future instances. Its aim to categorize data from prior information. It analyses and studies the training data and from that analysis it tries to infer a function which can be used for mapping new examples. In an optimal scenario, it will allow the algorithm to correctly determine the class labels for unseen instances. The learning algorithm should be able to generalize from the training data to future situations in a best plausible way.

Classification is used frequently in data science problems. Different techniques have been proposed to solve these problems viz. Rule-based techniques, Instance-based techniques, Logic-based techniques, stochastic techniques. Classification is essential to data analytics, pattern recognition and machine learning. It is a supervised learning technique, since it categorizes data from the prior information.

The class of each testing instance is decided by combining the features and finding patterns common to each class from the training data. Classification is done in two phases. First, a classification algorithm is applied on the training data set and then the extracted model is validated against a labeled test data set to measure the model performance and accuracy.

Applications of classification include document classification, spam filtering, image classification, fraud detection, churn analysis, risk analysis, etc. Few of the examples on supervised learning are Naive Bayes Classifier, decision trees, K-NN (nearest neighbors) algorithms etc.

2.1 Naïve Bayes Classifier

In naive bayes classifier, we have some the labelled datasets, form that we tried to trained the data, i.e. checkout the prior probability of the different class labels in the system. Prior probability is the probability which is taken priory, when the new data is not included. Not after taking care of prior probability we try to calculate posterior probability where we try to check the probability of that new data for which we have to derive class label, with the objects lying close to that object. After calculating both, we try to merge them and in the end from the majority of that, we try to find the which class label should be given to that newly arrived data object.

In order to know about Naive bays classifier more perfectly let us take a simple example best explains of Naive Bayes for classification. So, let's say we have data on 1300 people in a school. The people being students, teachers and the workers working in the school.

Table 1: Table on Naïve Bayes classifier

School	***Male***	***Female***	***Total***
Students	500	400	400
Teachers	100	50	150
Office workers	150	100	250
Total	750	550	1300

On the basis of our training set we can say the following:

- from 900 students 500 (0.56) are Male, 400 (0.44) are Female
- from 150 teachers 100 (0.67) are Male, 50 (0.33) are Female
- From 250 workers, 150 (0.6) are Male, 150 (0.4) are Female

Now suppose if a new person is introduced and we want to know the new person is either a student, teacher or a worker if the new person introduced is a male. We can predict the person being student teacher or a worker by the following formula. And we will check the probability of each of possibility and the one coming with the highest probability score will be our winner, and that class will be assigned to the new person introduced.

a) Student:

P(Student Male)=P(Student|Malestudent)*P(Student|Female student)*P(Student|Males)

=0.56*0.44*0.67

=.165088

b) Teacher:

P(Teacher|Male)=P(Teacher|Maleteacher)*P(Teacher|Female teacher)*P(Teacher|Males)

=0.67*.33*.133

=.0294063

c) Worker:

P(Worker|Male)=P(Worker|Maleworker)*P(Worker|Female worker)*P(Worker|Males)

=.6*.4*.2

=.048

After comparison of all the above probabilities, the probability of new male person being a student is very high so the new male person with get the class of student. based on the higher score (0.0165088).

2.2 Decision Trees

Decision trees is used to classify on the basis of decision, here a tree is constructed and has some different branches, these branches result to different class labels, if the label is satisfying the property of that class then the label is assigned to that input variable. The following could be an example for decision trees:

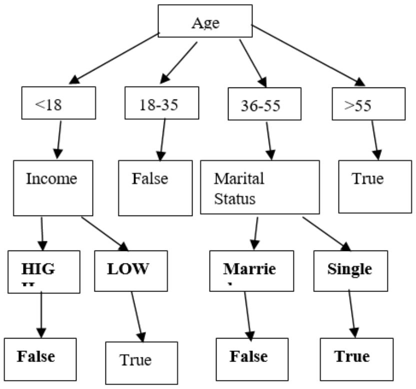

Fig. 2 An example on decision tress

Here also the different branches are all trained before, by the training set of inputs, so we get the labelled data. And these labelled will have different classes and these classes will help us in labelling the new input which comes for labelling.

2.3 K-NN Nearest neighbors

In K-NN nearest neighbor's algorithm, there are already defined class labels, used for training the machine. Then as the new input arises the new input is verified by checking its K nearest neighbors, the majority of the time K value is taken as odd, cause sometimes even values have equal distribution of class. So, it's hard to decide the class of different labels. In odd we can know about the majority of the class of its k nearest value. Majority of the class belonging to its K nearest neighbor will be given that class.

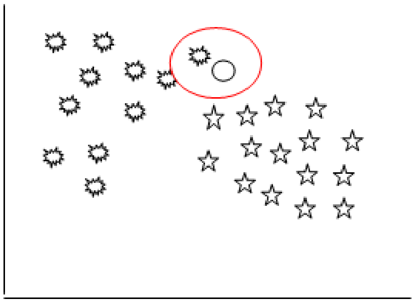

Fig 3: K-NN when K=1

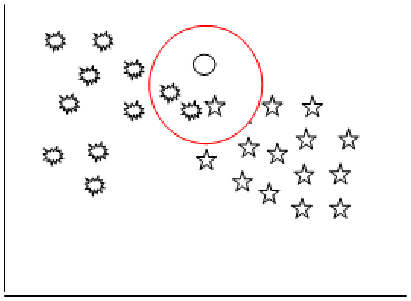

Fig 4: K-NN when K=3

3. Unsupervised Learning

In unsupervised learning, we don't have things like labelled data, or training data we just have some random data, and we try to group them by the behaviors of the data, if the data is showing some similarity then it will group by the behavior and similarities, in other words we are just trying to convert unlabeled data to labelled data. K means could be the best example for unsupervised learning

3.1 K-Means

K-means algorithm works similar to clustering, in clustering we try to combine the random data, into groups or we can say clusters, which are having similar properties similar behavior. In k-means algorithm we try to have k-centroids, the k centroid will be the number of classes which are needed to divide the data in, then we try to divide the different unlabeled data, into their nearest centroid labels, after that we try to estimate the distance where it should belong and then again, we try to change the place of centroid, and repeat that step again, until the centroid reach its proper place, and nearby data has been defined proper class to it.

4. Unsupervised Learning

In order to understand the meaning of reinforcement learning, we should know the meaning of reinforcement, Reinforcement means the result of strengthen the behavior such that it can perform better than how it was behaving before. And learning done on the basis of that, that is called reinforcement learning.

Markov decision process, Monte-Carlo approximation problem etc. are the examples of reinforcement learning.

4.1 Markov Decision Process

In Markova decision process, we have to keep the few things in mind, that is state which is it in, the action which is going to take and the state which it will come after doing that action. Markov in Markova decision process means only the present matters, that means it it's not concern about the things which you did in the past, or the state where you came from, it is just concern about the state where you and where you will end up. Then one more factor that is very important and which we should keep in mind is the reward. As in when you reach from one state to another, the action which you did in coming to the next state, it is beneficial for getting the result, or it is better than the other action if we take and end up on that state, i.e. the property of reward

4.2 MONTELCARLO Approximation

Its method of approximating things using samples, and mostly the things to approximate in machine learning are expectation, so an approximation of expectation leads to Monte Carlo approximation. Monte Carlo approximation, it relies on random experiment for obtaining numerical results. Monte Carlo was used by the scientist working on atom bomb, it is named after a place in Monaco, it is a small place which is famous for its gambling.

Now there are some data, whose behavior is little uncommon from the behavior of the class defined by the k centroid these are those items, which are having unusual behavior, these are called outlier or we can say anomaly

5. Outliers

The data items are those in which the observation is having very unusual behavior, those are called outliers. Outliers or anomaly are those which are deviated from the normal, or we can say the observation which are unexpected. There are few methods to detect outliers or anomalies these are as follows.

1) Statistical based approach
2) Depth Based approach
3) Distance based approach

5.1 Statistical based approach

In statistical based approach, we have a statistical distribution and we compute the statistical parameters of this statistical distribution such as the mean and the standard deviation. The outliers in this approach are those observations which have very less probability to lie in any kind of classes, even after lots of iteration. In simple words after lots of iteration of standard deviation from its mean, it still cannot be classified they become an outlier. Normal objects lie in this area but the objects we are strongly deviate from these observations, these are the outliers.

5.2 Depth based approach

the depth based approaches are those approaches, in which the observations are organized as some kind of convex hull layers. The class of the different observations will be decided on the basis of the depth, the observation in the similar class will have the same depth, whereas different classes will have different depth, and observation lying inside the convex hull layers are the observations which are normal, as outward the layer goes, then the chance of the observation to be an outlier increase, the observation which are lying in the outermost layer are those observations which can be called as an outlier. The convex hull approach or the depth based approach is only efficient if the observation is in 2D/3D.

5.3 Distance based approach

in the distance based approaches, we try to find out the distance of from its nearest neighbors, then slowly the cluster which is behaving to that observations are assigned that class, and the observations which are not even close to the neighbors are those observations which we can call as an outlier. Normally the observation which are normal will have the dense neighborhood, whereas the outlier will have lying a little far from those dense neighbored. One of the very good example for the distance based approaches is the K-NN method, in the KNN could be really good example for a distance based approaches.

5.4 Boxplot or whiskers plots

Box or whisker plots is a detection technique for outlier detection, it's a graphical method for identifying outliers. Where you have some observation, you plot those observations in a graph and then we try to identify outliers from these observations. These observations have a certain limit we try to plot a box from these observation, and the observation which are not lying inside this box we call them as an outlier.

Let's understand it by an example: -
Following are few observation
2,51,53,54, 43,51,62,49,50, 63,60.

Arrange those in ascending order
2,43,49,50,51,51,53,54,60,62, 63.

we find the position of the median

position of median = 11+1/2 = 6th
smallest observation = 2
lower quartile: - The median of the lower half of the observation
lower quartile = 49
median = 51 (6th observation)
upper quartile: - The median of the upper half of the observation
upper quartile = 60
highest observation = 63

now we try to calculate the inner quartile range i.e. the difference of upper quartile value with its lower quartile value = 60-49=11
IQR (inner quartile range) =11

Now let's try to define the range of the observation and we try to know which values are outliers from the above observation

Lower quartile – 1.5(IQR) = 49 – 1.5(11)
= 32.5
Upper quartile +1.5(IQR) = 49 +1.5(11)
= 76.5

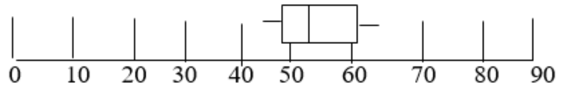

Fig 4: Boxplot and whisker plots for outlier detection

References

[1] Junfei Qiu, Qihui Wu, Guoru Ding, Yuhua Xu and Shuo Feng (2016). A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing Santosh Kumar,Xiaoying Gao, Ian Welch (March 2016) A Machine Learning Based Web Spam Filtering Approach on Advanced Information Networking and Applications (AINA), 2016 IEEE 30th International Conference.

[2] Vanita Jain, Arun Dubey, Amit Gupta, and Sanchit Sharma (March 2016) Comparative analysis of machine learning algorithms in OCR on Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference

[3] Payal A. Jadhav, Prashant N. Chatur, Kishor P. Wagh (February 2016) Integrating performance of web search engine with Machine Learning approach on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2016 2nd International Conference.

[4] T T Dhivyaprabha, P Subashini, M Krishnaveni (December 2016) Computational intelligence based machine

learning methods for rule-based reasoning in computer vision applications in Computational Intelligence (SSCI), 2016 IEEE Symposium Series.

[5] TM Mitchell, Machine learning (McGraw-Hill, New York, 1997)

[6] S Russell, P Norvig, Artificial intelligence: a modern approach (Prentice- Hall,EnglewoodCliffs, 1995)

[7] V Cherkassky, FM Mulier, Learning from data: concepts, theory, and methods (John Wiley & Sons, New Jersey, 2007)

[8] TM Mitchell, The discipline of machine learning (Carnegie Mellon University,

[9] School of Computer Science, Machine Learning Department, 2006)

[10] C Rudin, KL Wagstaff, Machine learning for science and society. Mach Learn 95(1), 1–9 (2014)

[11] CM Bishop, Pattern recognition and machine learning (Springer, New York, 2006)

[12] N Jones, Computer science: the learning machines. Nature 505(7482), 146–148 (2014)

[13] J Langford, Tutorial on practical prediction theory for classification. J Mach Learn Res 6(3), 273–306 (2005)

[14] R Bekkerman, EY Ran, N Tishby, Y Winter, Distributional word clusters vs. words for text categorization. J Mach Learn Res 3, 1183–1208 (2003)

[15] R Bekkerman, EY Ran, N Tishby, Y Winter, Distributional word clusters vs. words for text categorization. J Mach Learn Res 3, 1183–1208 (2003)

Dr. Akhtar Rasool received the B.E. M.Tech. and PhD degrees in Computer Science Engineering. He has a specialization in string matching algorithms, parallel computing, Artificial intelligence, Big Data Analysis, Cluster and Grid computing. He has published around 10 papers in international journals and attend an international conference. He is currently an assistant professor in Maulana Azad National Institute of Technology.

Zeeshan Ahmad Lodhia received the B.Tech. in Information Technology. In Dharamsinh Desai University of Nadiad Gujarat. from 2010-2014. Currently he is now pursuing M.Tech in Computer Science Engineering in Maulana Azad National Institute of Technology, under the guidance of Dr. Akhtar Rasool.

Gaurav Hajela is pursuing Ph.D from computer science department of Maulana Azad National institute of technology, and has done M.tech with specialization in computer networks from MANIT, Bhopal.