JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

MACHININE LEARING IN NARRATIVE INTELLIGENCE

Santhanam, Assistant Professor, Department of Information Technology, Dhanalakshmi Srinivasan College of Engineering and Technology

K. Poornima Devi,Assistant Professor, Dhanalakshmi Srinivasan College of Engineering and Technology

ABSTRCT:

In this project, we were asked to experiment with a real world dataset, and to explore howmachine learning algorithms can be used to find the patterns in data. We were expected to gainexperienceusingacommondata-miningandmachinelearninglibrary,Weka,andwereexpectedto submit a report about the dataset and the algorithms used. After performing the required tasksonadatasetofmychoice,hereinliesmyfinalreport. Machine Learning can be thought of as the study of a list of sub-problems, viz: decisionmaking, clustering, classification, forecasting, deep-learning, inductive logic programming,support vector machines, reinforcement learning, similarity and metric learning, geneticalgorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machinelearning task of inferring a function from a labeled data .

Keywords: MachineLearning,PatternRecognition,Classification,Supervisedlearning,ArtificialIntelligence

INTRODUCTION:

Machinelearningisasub-domainofcomputersciencewhichevolvedfromthestudyofpattern recognition in data, and also from the computational learning theory in artificialintelligence. It is the first-class ticket to most interesting careers in data analytics today[1]. Asdata sources proliferate along with the computing power to process them, going straight to thedataisoneofthemoststraightforwardwaystoquicklygaininsightsandmakepredictions.

Machine Learning can be thought of as the study of a list of sub-problems, viz: decisionmaking, clustering, classification, forecasting, deep-learning, inductive logic programming,support vector machines, reinforcement learning, similarity and metric learning, geneticalgorithms, sparse dictionary learning, etc. Supervised learning, or classification is the machinelearning task of inferring a function from a labeled data [2].

In Supervised learning, we have atraining set, and a test set. The training and test set consists of a set of examples consisting ofinput and output vectors, and the goal of the

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

supervised learning algorithm is to infer a functionthat maps the input vector to the output vector with minimal error. In an optimal scenario, amodel trained on a set of examples will classify an unseen example in a correct fashion, whichrequires the model to generalize from the training set in a reasonable way. In layman's terms,supervisedlearningcan betermedasthe processofconceptlearning, whereabrainis exposedtoa set of inputs and result vectors and the brain learns the concept that relates said inputs tooutputs. A wide array of supervised machine learning algorithms are available to the machinelearning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines,RandomForest,NaïveBayesClassifier, Bayes Net,MajorityClassifier[4,7,8,9] etc.,andthe each have their own merits and demerits. There is no single algorithm that works for all cases, asmeritedbythe Nofreelunchtheorem[3]. Inthisproject,wetryandfindpatternsinadataset[2],which is a sample of males in a heart-disease high risk region of South Africa, and attempt tothrowvariousintelligently-pickedalgorithmsatthedata,andseewhatsticks.

PREVIOUS WORK:

The procedure described in this contribution is based on non-adaptive meta-models. This means, the sample set is pre-computed and the meta-model is trained on thebasis of this data. Adaptive meta-models integrate the sampling and training in the usagephase [Dubourg (2011)]. Due to the adaptivity, they perform better for highly non-linearproblemsbutaremorecomplicatedinimplementationandusage.

DISADVANTAGES

- Massive data sets to train on, and these should be inclusive/unbiased, and of good quality
- ML needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy
- Challenge is the ability . To accurately interpret results generated by the algorithms.

PROPOSED WORK:

To experiment with a real world dataset, and to explore howmachine learning algorithms can be used to find the patterns in data. We were expected to gainexperienceusingacommondata-miningandmachinelearninglibrary,Weka,andwereexpectedto submit a report about the dataset and the algorithms used.

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

ADVANTAGES

- The ability to learn, it lets them make predictions and also improve the algorithms on their own.
- Can review large volumes of data and discover specific trends and patterns that would not be apparent to humans.
- Gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions.
- Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

4. MODEL OF EXPERIMENTS

4.1 PROBLEMSAND ISSUESINSUPERVISEDLEARNING: Before we get started, we must know about how to pick a good machine learningalgorithmforthe givendataset.Tointelligentlypickanalgorithmtouseforasupervisedlearningtask,wemustconsider thefollowingfactors[4]:

HeterogeneityofData:

Many algorithms like neural networks and support vector machines like theirfeature vectors to be homogeneous numeric and normalized. The algorithms thatemploy distance metrics are very sensitive to this, and hence if the data isheterogeneous,thesemethodsshouldbetheafterthought.DecisionTreescanhandleheterogeneousdataveryeasily.

RedundancyofData:

Ifthedatacontainsredundant information,i.e.containhighlycorrelatedvalues,then it's useless to use distance based methods because of numerical instability. Inthis case, some sort of Regularization can be employed to the data to prevent thissituation.

DependentFeatures:

If there is some dependence between the feature vectors, then algorithms thatmonitorcomplexinteractionslikeNeuralNetworksandDecisionTreesfarebetterthanotheralgorithms.

Bias-VarianceTradeoff:

A learning algorithm is biased for a particular input x if, when trained on each ofthese data sets, it is systematically incorrect when predicting the correct output for x,whereas a learning algorithm has high variance for a particular input x if it predictsdifferent output values when

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

trained on different training sets. The prediction error ofa learned classifier can be related to the sum of bias and variance of the learningalgorithm,andneithercanbehighastheywillmakethepredictionerrortobehigh.Akey feature of machine learning algorithms is that they are able to tune the balancebetween bias and variance automatically, or by manual tuning using bias parameters,andusingsuchalgorithmswillresolvethissituation.

CurseofDimensionality:

If the problem has an input space that has a large number of dimensions, and theproblem only depends on a subspace of the input space with small dimensions, themachine learning algorithm can be confused by the huge number of dimensions andhence the variance of the algorithm can be high. In practice, if the data scientist canmanually remove irrelevant features from the input data, this is likely to improve theaccuracy of the learned function. In addition, there are many algorithms for featureselectionthatseektoidentifytherelevantfeaturesanddiscardtheirrelevantones,forinstance **Principle Component Analysis** for unsupervised learning. This reduces thedimensionality.

Overfitting:

Theprogrammershouldknowthatthereisapossibilitythattheoutputvaluesmayconstitute ofaninherentnoisewhichistheresultofhumanorsensorerrors.Inthis case, the algorithm must not attempt to infer the function that exactly matches all thedata. Being too careful in fitting the data can cause overfitting, after which the modelwill answer perfectly for all training examples but will have a very high error forunseen samples.

4.2 DATASET:

The dataset used is a sample of males in a heart-disease high-risk region of the WesternCape,SouthAfrica.Thedatasetthatwasusedforthisprojectisasubsetofamuchlargerd ataset,as described inRousseauw et al, 1983, South African Medical Journal, and has the followingfeaturevectors:

- Sbp - systolicbloodpressure
- Tobacco - cumulativetobacco(kg)
- Ldl - lowdensitylipoproteincholesterol
- Adiposity - thisistheamountoffattissueinthebody
- Famhist - familyhistoryofheartdisease(Present,Absent)(String)
- Typea - type-Abehavior
- Obesity - Stateofbeingoverweight

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

- Alcohol - currentalcoholconsumption
- Age - Ageatonset
- Chd - coronaryheartdisease(ClassLabeloftheDataset)

In the dataset, there are 462 example vectors. Expert Systems have been used in the fieldof medical science to assist the doctors in making certain diagnoses, and this can help save lives.Coronary Heart Disease is a disease where a waxy substance builds up inside the coronaryarteries,andhencethismay leadtoheartattack,andeven death.Whendiagnosed andtreated,the treatment can go a long way in helping the patient. This classification task is importantbecause the expert system, when correctly generalized, can tell the doctor which patient mayhave the disease, and the doctor can take a look at that case in more detail. Moreover, if thedoctor makes a slip, i.e. misdiagnoses someone, the expert system can help rectify his mistake. Itresultsintwodoctors,oneofthemvirtual,insteadofonedoctordiagnosingeverycase whichhasagreaterchanceofaccuracyandprecision.
Firstweperformthesignificanceanalysisofthe9featurevectors,toseewhichvectorshave more significance in representing the classes.

We used **Principal ComponentAnalysis**[4,7,9] forthispurposeandcameupwiththefollowingresults.
AttributeEvaluator(supervised,Class(nominal):10chd):

0.6795	10.516adiposity+0.46age+0.401 ity+0.334ldl+0.324sbp...
0.5465	20.543alcohol+0.459tobacco- 0.392obesity-0.364ldl-0.282typ
0.4269	3-0.792typea- 0.459alcohol+0.338famhist+0.1 e+0.125sbp...
0.322	4-0.833famhist-0.305obesity- 0.258alcohol-0.21typea-0.196sb

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

Correlationmatrix

1 0.21 0.160.36 -0.09 -0.060.24 0.140.39

0.211 0.160.29-0.09 -0.01 0.120.2 0.45

0.160.161 0.44-0.160.040.33-0.030.31

0.360.290.441 -0.18-0.040.720.1 0.63

-0.09-0.09-0.16-0.181 -0.04-0.12 -0.08 -0.24

Rankedattributes:

0.2291 50.624tobacco-0.419alcohol+0.321typea+0.305hist-0.283obesity...

0.1446 60.781sbp-0.379alcohol+0.332typea-0.215 0.174obesity...

0.0706 70.788ldl-0.333obesity+0.277alcohol+0.2p-0.196adiposity...

0.0194 80.691age-0.489tobacco-0.339obesity-0.235sbp+0.187famhist...

Selectedattributes: 1,2,3,4,5,6,7,8:8

Here we can see that all factors are important after we do the PCA. The last feature hasbeen deemed unworthy by the PCA implementation in WEKA, which made little sense to us asageis highlycorrelatedtomost diseases.Wefurtherourinvestigationbyusinganotherattributeselector, the **SignificanceAttributeEvaluator**[5,9], onthedatatoyield

Significanc featureevaluatorRankedattributes:

0.3019age

0.2992tobacco

0.2936typea

Selectedattributes: 9,2,6,5,3,4,1,7,8:9

Here, we see that feature 9, i.e. the age of the patient was the most significant factor forclassification purposes, and factors 7 and 8, obesity and alcohol consumption were the leastsignificantfactors.ThroughcombinedresultsofPCAandSAE,weconcludethatallthefeatures

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

are relevant for our purposes. The name of the sample was removed as well. Except for the useofPCAandSAE,nootherpre-processingwasdoneonthedata.

BaselineClassifier:

As the baseline classifier, we chose a Naïve Bayesian Network because it is easy tocompute, and because the features in the given dataset are all aspects of a person's physicalhabits or medical history, and hence can be assumed to be independent of each other, which isthe primary assumption in Naïve Bayes Classifier[6,8,9]. It is a conditional probability model:given aprobleminstance tobeclassified,representedbyavector representingsomenfeatures(independentvariables),itassignstothisinstanceprobabilitiesforeach ofKpossibleoutcomesorclasses.Theproblemwiththeaboveformulationisthatifthenumberoffeat uresnislargeorifafeaturecantakeonalargenumberofvalues,thenbasinH

suchamodelonprobabilitytablesisinfeasible.Wethereforereformulatethemodeltomakeitmoret ractable.UsingBayes'theorem,theconditionalprobabilitycanbedecomposedas

$$p(C_k|\mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

So, assuming that $p(x_i|x_{i+1}, \ldots, x_n, C_k) = p(x_i|C_k)$ for i = 1, ….. n-1. So, under theindependenceassumptions,wecansaythat Where,theevidence isascalingfactordependentonlyon ,thatis,aconstantifthevaluesofthefeaturevariables areknown.

Untilnow,wehavederivedanindependentfeaturemodel.InNaïveBayesclassifier,wecombine this model with a decision rule, and one of the common rules is to pick whichhypothesis is the most probable. The corresponding classifier, a Bayes classifier, is the functionthatassignsaclasslabel $\hat{y} = C_k$ forsomekasfollows:

The Naïve Bayes network is already implemented in the Machine Learning libraryWEKA[5,9].Thiswasusedontheaforementioned dataset,whichledtothefollowingoutput: As we can see that the Naïve Bayes classifier works really well with the given dataset,with the True Positive classification rate being 71.6 percent on an average, i.e. this classifier cancorrectly classify 71.6 percent of all the examples it sees. However, there is still a vast majorityof the dataset, i.e. 28.4% which can't be correctly classified. This means that our expert medicaldiagnosis system stillmisdiagnoses onethirdofits cases,andonethird

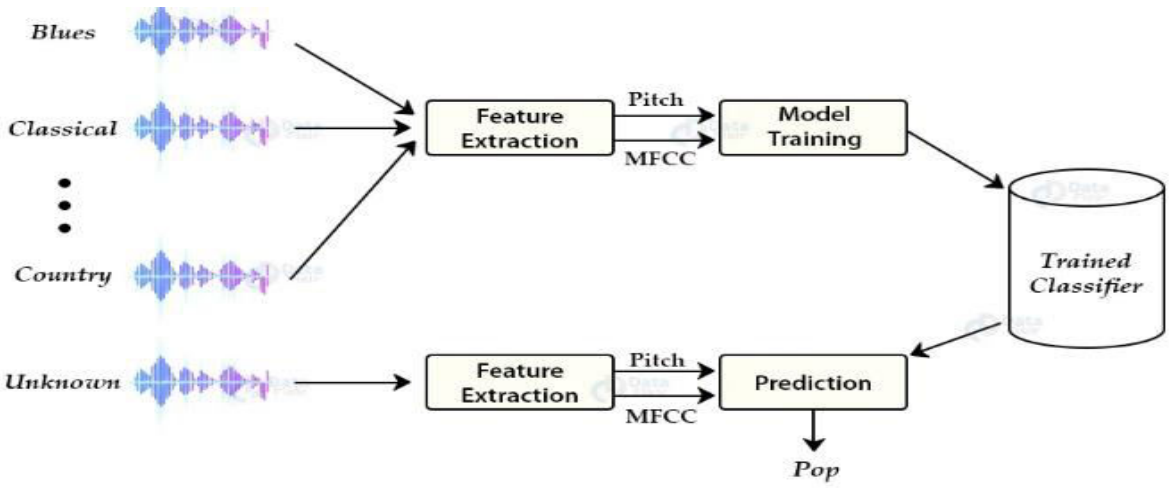
ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

ofthepatients'symptom whomayhavethediseasearenotbeingscrutinizedbythedoctor.Wewill nowattempttoimproveresultbyusingothermoresophisticatedclassifierS.

4.3 CLASSIFIER:

Since it's a binary dataset with the class label being either the person has CHD or s/hedoesn't have CHD, and the number of samples is less than 100 times the number of features, thecorrelation matrix shows us that the correlation between various features is under . 5, we believethat support vector machines would be a viable classifier in this case.

SYSTEM ARCHITECTURE:



OUTPUT:

===DetailedAccuracy ByClass===

TPRateFPRatePrecisionRecallF-MeasureMCC ROCAreaPRCArea Class

0.825 0.506 0.755 0.825 0.788 0.335 0.659 0.737 0

0.494 0.175 0.598 0.494 0.541 0.335 0.659 0.471 1

WeightedAvg. 0.710 0.392 0.700 0.710 0.702 0.335 0.659 0.645

===ConfusionMatrix ==

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

ab<-- classified as

24953|a=0

8179|b=1

Here, we can see that the said SVM performs better than the Naïve Bayes classifier forclass 0, predicting 82.5% of the classes correctly, whereas it performs slightly worse than NaïveBayes for class 1 with 49.4%. On an average, the true positive rate was achieved to be 71% ascompared to 71.6% in case of Naïve Bayes. This result is surprising, as we expected SVM toperformbetterthantheNaïveBayesClassifierfor independentnon-redundantfeaturevectorsasSVM projects low-dimensional sub-space to a higher dimensional subspace where the featuresare linearly separable. The RMS error for SVM was comparatively higher compared to NaïveBayes by .10 and the kappa statistic of Naïve Bayes was lower than SVM by .05, which showsthatNaïveBayesisthebetterclassifier. Curious about why the data was behaving the way it was, we did use other classifiers onthe said dataset . We used Multilayer Perceptron, Decision Tree (J48) [8,9], Random Forest[8,9]with 100 trees, and the only classifier that got close was the J48 with true positive rate of 70.7%.performed poorly with only 63% TPR, -

learningneuralnetperformedwith65.38%correct classifications.CuriousifLazylearning[8,9]could do any better, we tried it and found that it correctly classified 61.25% of the cases. Theonly thing we could now think of is that the input space was incomplete, and needed moredimensionsforbetterpredictions,andwiththegivenfeaturevectors,

CONCLUSION:

We conclude that the dataset is not a complete space, and there are still other featurevectors missing from it. What we were attempting to generalize is a subspace of the actual inputspace, where the other dimensions are not known, and hence none of the classifiers were able todobetterthan71.6%(NaïveBayes).Inthefuture,ifsimilarstudiesareconductedto generatethedataset used in this report, more feature vectors need to be calculated so that the classifiers canformabetterideaoftheproblemathand.

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

FUTURE WORK:

The prospect of Machine Learning is not limited to the investment sector. As the Machine Learning prospect is very high, there are some of the areas where researchers are working toward revolutionizing the world for the future. Thus, the future prospect of Machine Learning will accelerate the processing power of the automation system used in various types of different collection technologies.

Further, in this blog on the future prospect of Machine Learning, we will look into the skills that are required to become an ML Engineer. In this blog on the future prospect of Machine Learning, we have looked around the need for Machine Learning. We have seen the future prospect of Machine Learning and the opportunities in the field.

REFERENCES:

- ElementsofStatisticalLearning:DataMining,Inference,andPrediction. 2ndEdition.
- Datasets:CoronaryHeartDiseaseDataset."ElementsofStatisticalLearning:DataMining, Inference, and Prediction. 2nd Edition. Accessed April 27, 2016.http://statweb.stanford.edu/~tibs/ElemStatLearn/.
- "NoFreeLunchTheorems."No FreeLunchTheorems.Accessed April27, 2016.http://www.no-free-lunch.org/.
- Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. The Elements of StatisticalLearning:DataMining,Inference,andPrediction:With200Full-colorIllustrations.NewYork:Springer,2001.
- "Weka 3: Data MiningSoftware in Java."Weka 3.AccessedApril27,2016.http://www.cs.waikato.ac.nz/ml/weka/.
- Bozhinova,Monika,NikolaGuid,andDamjanStrnad.Naivni Bayesov Klasifikator:DiplomskoDelo.Maribor:M.Bozhinova,2015.
- Schölkopf, Bernhard, ChristopherJ. C. Burges, andAlexanderJ. Smola.Advances inKernelMethods:SupportVectorLearning.Cambridge,MA:MITPress,1999.

ISSN- 2394-5125 VOL 6, ISSUE 06, 2019

Norving, Peter,and StuartRussel.ArtificialIntelligence:AModernApproach.S.l.:PearsonEducationLimited,2013.

- Witten,I.H.,andEibeFrank.DataMining:PracticalMachineLearningToolsandTechniques.Amsterdam:MorganKaufman,2005.
"IntrotoMachineLearning|Udacity."IntrotoMachineLearning|Udacity.AccessedApril27,2016.https://www.udacity.com/course/intro-to-machine-learning--ud120.