The University of Texas at Dallas

Project Final Report

Machine Learning [6375.501]

Analysis of 2016KDD_CUP Dataset using ML Techniques

Submitted By

Nidhi Jain [nxj150930]

April 24th 2016

Introduction

Machine Learning techniques are important and useful in many data processing fields. Finding influential attributes from given dataset for identifying patterns and thus making a strong prediction for future instances is becoming a highly valued topic in today's research industry.

This project applies some similar kind of Machine Learning techniques on "2016_KDD_CUP" dataset to analyze the selection pattern of research papers in various conferences influenced by certain attributes. Next section will provide a little background of the KDD dataset, attributes, instances & problem statement.

Problem & Data Description

For students, parents and funding agencies that are planning their academic pursuits or evaluating grant proposals, having an objective picture of the institutions in question is particularly essential. The KDD Cup dataset is a publicly and freely available dataset that includes information of academic publications and citations. This dataset can be used to study the influential nodes of type "affiliations (institutions)" and "conference locations".

In effect, given a research field, we are challenging the Machine Learning techniques to get trained on the given dataset & identify/rank the best research institutions based on their publications. Another innovative and interesting task is: given any instance of upcoming top conferences such as KDD, SIGIR, and ICML in 2016, rank the importance of institutions based on predicting how many of their papers will be accepted.

KDD Dataset format on Internet

KDD dataset is provided on internet as a combination of following text files:
[Papers, Affiliations, Conferences, Field of Study, Authors, Selected Papers, Selected Affiliations]

KDD Dataset format Transformation

To effectively apply the classifying and prediction techniques of Machine Learning on KDD dataset, I transformed the dataset to single (.csv) comma delimited file named "KDD_Transformed_Dataset.csv" including only influential attributes.

Attributes of data file (KDD_Transformed_Dataset.csv)

- Publish_Date //Year when Paper published
- Field_Id //Id of "Field of Study" that Paper belongs to
- Field_Name //Name of "Field of Study" that Paper belongs to
- Author_Id //Id of Author who wrote the Paper
- Author_Name //Name of Author who wrote the Paper
- Affiliation_Id //Id of Affiliation/Institute that Paper belongs to
- Affiliation_Name //Name of Affiliation/Institute that Paper belongs to
- Conf_Name //Conference name in which Paper is presented

- Location_Id //Conference Location Id
- Conference_Location //Location Name where Conference held
- Paper_Selected //Class: 1 = Paper & Affiliation selected in Conference
  0 = Paper & Affiliation not selected in Conference

Number of Instances in KDD_Transformed_Dataset.csv file: 358961

Experimental Methodology

Firstly, the project trains 7 classifiers () on given KDD dataset. The accuracy of classifiers is mentioned in Data_Analysis_Report.doc. Then the whole dataset is predicted using those classifier models to generate data from classifier's point of view. The resulting predictions are converted into .csv files for each classifier.

After having classier specific predictions in hand, following 5 experiments are conducted on each of those predicted .csv datasets and the matching results from all classifiers are considered as the conclusion. 5 Experiments conducted in predicted data of each classifier are as follows:

- The first experiment analyzed the influence of attributes (field study, conference location, author, affiliation) on Paper's selection and hence generated the rank of research institutions/affiliations (top 10) presenting such papers in (last 3 years).

- The second experiment analyzed the top 10 selected institutes/affiliations in some specific conferences (KDD, MM, MOBICOM, SIGCOMM, SIGIR, SIGMOD) in last 3 years.

- The third experiment analyzed the top 10 field of study whose papers are mostly selected in some specific conferences (KDD, MM, MOBICOM, SIGCOMM, SIGIR, SIGMOD) in last 3 years.

- The fourth experiment analyzed the top 10 authors whose papers are mostly selected in some specific conferences (KDD, MM, MOBICOM, SIGCOMM, SIGIR, SIGMOD) in last 3 years.

- The fifth experiment analyzed the impact of conference location on total no. of Papers presented by different affiliations/institutions & selected in some specific conferences (KDD, MM, MOBICOM, SIGCOMM, SIGIR, SIGMOD) in last 3 years.

Classifiers used

- Decision Tree
- Naïve Bayes
- Logistic Regression
- Random Forest
- SVM [Gaussian]
- Bagging
- Boosting

Training the Classifiers

The training part of Classifiers depends on the type of learning technique used. In this project the learning is "Supervised" and thus the training part identifies a function/hypothesis from labeled data which can be used for mapping new examples. The training data consist of a set of training examples/instances [80% sampled over complete dataset] where each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal or labeled class). The remaining 20% sample of dataset will later be used towards testing part.

The Classifiers are trained by calling the corresponding functions from packages and passing the attributes as required parameters as below:

- ***dtree_model <- fit(class_attribute ~ attribute2 + attribute3…, data=train, model="rpart", task='class')***
  //for training the Decision Tree Classifier

- ***svm_model <- fit(class_attribute ~ attribute2 + attribute3…, data=train, model="ksvm")***
  //for training SVM [Gaussian] Classifier

- ***nbayes_model <- fit(class_attribute ~ attribute2 + attribute3…, data=train, model="naiveBayes")***
  //for training Naïve Bayes Classifier

- ***rforest_model <- fit(class_attribute ~ attribute2 + attribute3…, data=train, model="randomForest")***
  //for training Random Forest Classifier

- ***lregression_model <- fit(class_attribute ~ attribute2 + attribute3…, data=train, model="lr")***
  //for training Logistic Regression Classifier

- ***bagging_model <- fit(class_attribute ~ attribute2 + attribute3…, data=train, model="bagging")***
  //for training Bagging Classifier

- ***boosting_model <- fit(class_attribute ~ attribute2 + attribute3…, data=train, model="boosting")***
  //for training Boosting Classifier

Prevention of Over-Fitting

Over-fitting is a term used to describe too much learning by any Classifier such that the training error becomes way too less but test error increases. This happens because the Classifier when spends too much time to learn each instance and concludes result for them independently, it becomes able to classify each training instance correctly but fails to classify the test data as the model is not generalized. To avoid this over-fitting I am going to apply following techniques for KDD dataset training part:

- ***Cross-Validation:*** Cross-validation separates model selection (training data) from testing (test data), resulting in a more conservative estimate of generalization. In this project I have sampled the KDD dataset as "80% training data" and "20% test data"

- ***Obtaining more training data:*** The more the data (instances), the less are the chances of model over-fitting as it becomes bound to generalize the model due to abundance of data.

In this project I have used abundance of training data containing more than 3,00,000 instances.

- ***Regularization:*** Regularization controls the penalty for complexity, which (when successful) will prevent under- and over-fitting. Many machine learning algorithms come with a knob that controls over-fitting. Typically in many algorithms we minimize some linear combination of the error on our training set and a regularization term, and there is a knob that trades off between these two terms. The regularization term, such as the squared norm of the weight vector in an SVM, penalizes the complexity of the learned model and favors simpler models. Too high a weight on the regularization and the model under-fits. Too low a weight and the model over-fits.

- ***Reducing Parameters:*** For algorithms that don't have a regularization term, such as decision trees, reducing the number of parameters/attributes prevents over-fitting. In Decision Trees reducing the attributes reduces the depth of a decision tree, thus leading to more generalized tree rather than having too many branches for each instance. In this project I have reduced the attributes of KDD dataset to some specific influential parameters (Paper's_Field_of_Study, Author, Affiliation, Conference_Location, etc) and thus removed many noisy and extra attributes (Paper_Title, Conference_Url, etc).

Prevention of null or redundant training data

If the input features contain redundant information (e.g., highly correlated features), some learning algorithms (e.g., linear regression, logistic regression) will perform poorly because of numerical instabilities. These problems can often be solved by imposing some form of regularization like assigning weights. In this project I am removing the instances that contain null data in influential attributes to prevent any instability in results. I am checking for missing values by using sapply function as follows:

- ***sapply (training_data.raw, function(x) sum(is.na(x)))***

Testing the Classifiers

In a real-world application of supervised learning, we have a training set of examples with labels, and a test set of examples with unknown labels. The whole point is to make predictions for the test examples. However, in research or experimentation we want to measure the performance achieved by a learning algorithm and Classifiers. To do this we use a test set consisting of examples with known labels. We train the classifier on the training set, apply it to the test set, and then measure performance by comparing the predicted labels with the true labels (which were not available to the training algorithm). In this project I am using 30% of KDD dataset sample as test data.

Prediction of test data class is done in following manner:

- ***predicted_output <- predict (trained_model, test_data)***

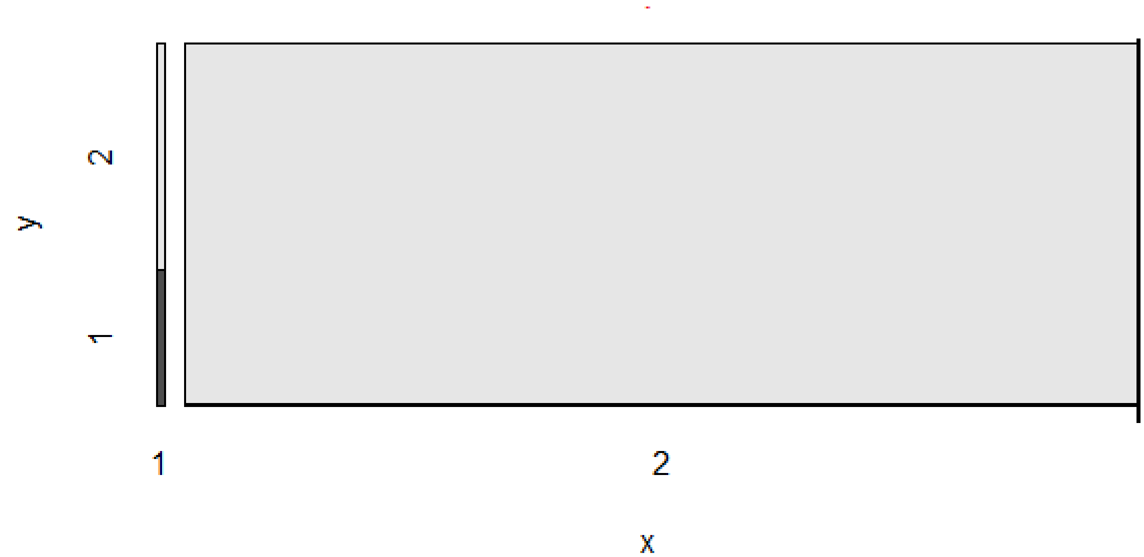
Visualization of results

Result of Classifier model predictions can be visualized by plotting graphs and charts using various functions available in R programming. In this project I am using following plot function:

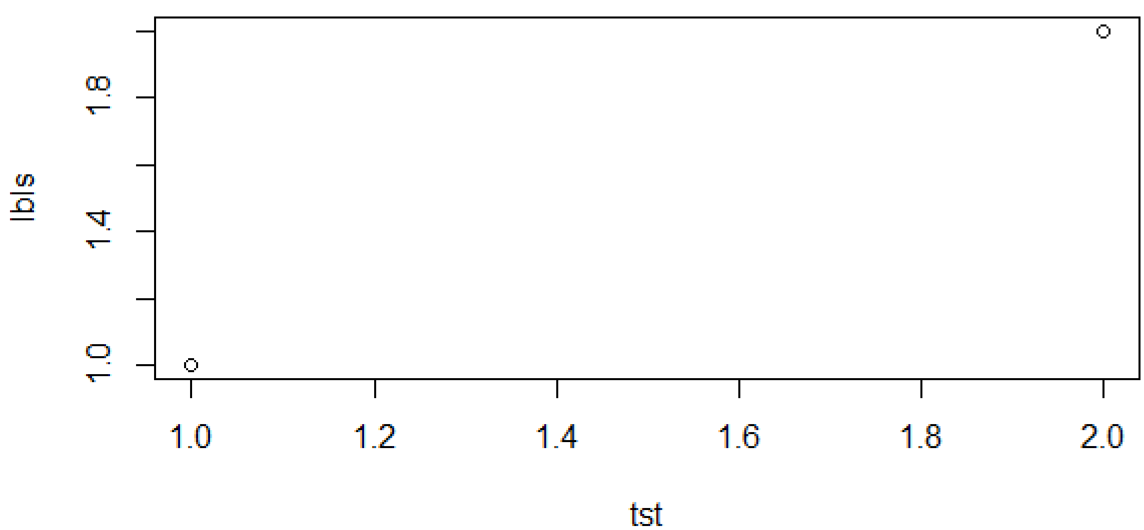
- ***plot (actual class of test data, predicted class of test data)***

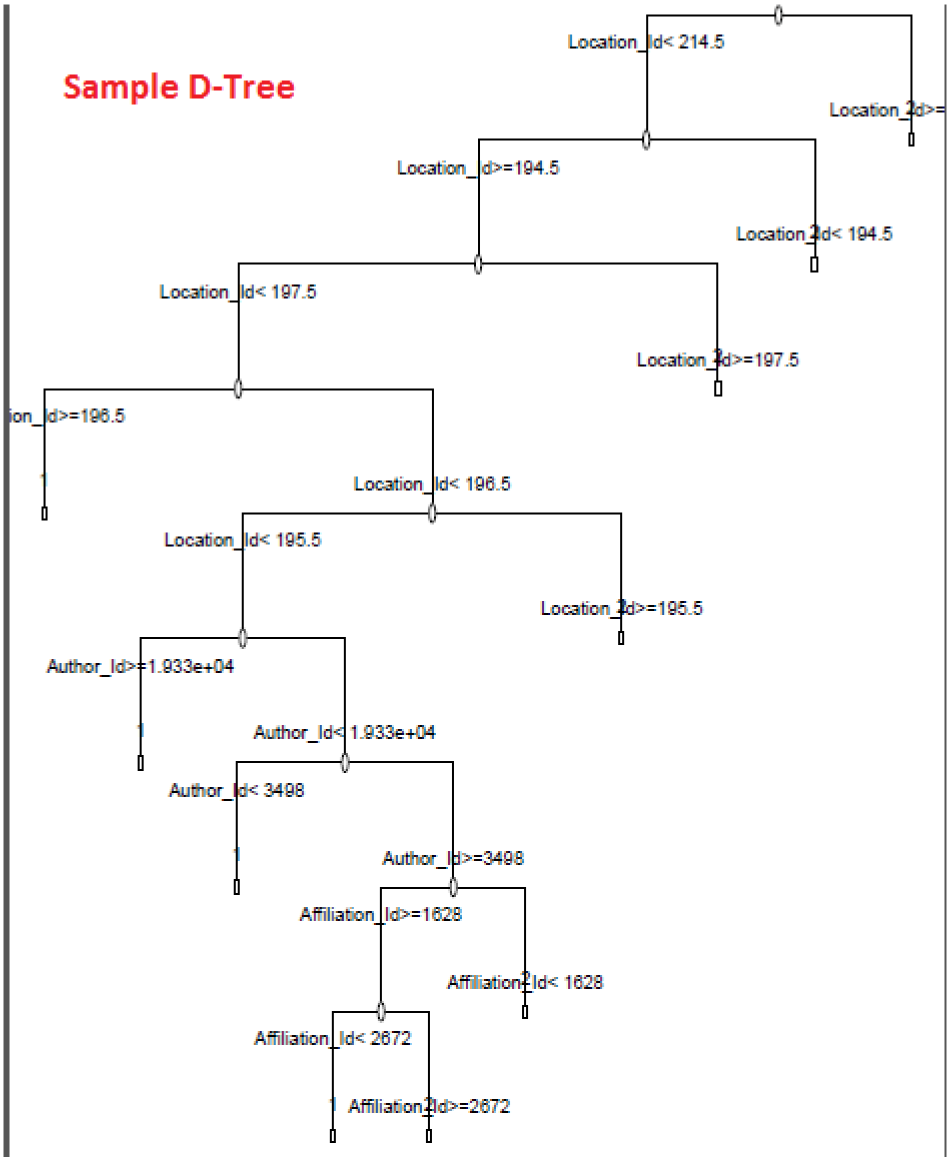
As our dataset contains only two discrete classes (1 & 2) the plot looks like below. Class 1 instances are way too less compared to class 2 instances, the plot looks like following graph containing less of class 1 instances than class 2.

D-Tree Plot (Actual vs Predicted)



R-Forest Plot (Actual vs Predicted)





Validation of results

Validation of results in nothing but analyzing the performance of trained_classifier_model. In this project I used following factors for performance analysis:

- Accuracy(in %): [(TP + TN) / (TP + TN + FP + FN)] * 100
- Precision: (TP) / (TP + FP)
- Recall: (TP) / (TP + FN)
- F-measure: (2TP) / (2TP + FP + FN)

Where; TP = True prediction of Positive Label (Actual 1, Predicted 1)
TN = True prediction of Negative Label (Actual 0, Predicted 0)
FP = False prediction of Positive Label (Actual 0, Predicted 1)
FN = False prediction of Negative Label (Actual 1, Predicted 0)

Coding Techniques

Language: R programming (for training Classifiers)

Platform/Tool: R Studio(for training Classifiers)
SQL server management studio (to run experiments on predicted data)

R Packages:

- rminer //for Classifiers
- rpart //used by rminer for Decision Tree
- kernlab //used by rminer for SVM
- e1071 //used by rminer for Naïve Bayes
- randomForest //used by rminer for randomForest
- ROCR //used for performance parameters (precision, recall, f-measure)
- stats //used by rminer for Logical Regression
- gplots //used by above packages

Classifiers Performance Report

Result from training classifiers using R code

Classifier	Precision	Recall	F-measure	Accuracy(%)
D-Tree	0.9952	0.9984	0.9968	99.3648
N-Bayes	0.9923	1	0.9961	99.2311
Logical-Reg	0.9923	1	0.9961	99.2311
R-Forest	0.993	0.9993	0.9961	99.234
SVM	0.9916	1	0.9958	99.1643
Bagging	1	1	1	100
Boosting	0.993	0.9986	0.9958	99.1643

Experiment Results

Experiment 1 Result

Top 10 affiliations with maximum number of Papers Selected

	Affiliation_Name	No_of_Papers_Selectd
1	microsoft	143
2	tsinghua university	102
3	google	102
4	ibm	94
5	massachusetts institute of technology	82
6	carnegie mellon university	79
7	national university of singapore	75
8	nanyang technological university	65
9	university of california berkeley	60
10	university of wisconsin madison	59

Top 10 Affiliations and Universities

Experiment 2 Result

Top 10 affiliations with maximum number of Papers Selected in Various Conferences

KDD	MM	MOBICOM
Affiliation_Name	Affiliation_Name	Affiliation_Name
ibm	national university of singapore	massachusetts institute of technology
carnegie mellon university	google	university of massachusetts amherst
arizona state university	microsoft	tsinghua university
university of southern california	tsinghua university	microsoft
ludwig maximilian university of munich	carnegie mellon university	university of wisconsin madison
google	chinese academy of sciences	nanyang technological university
tsinghua university	goldsmiths university of london	university of illinois at urbana champaign
hong kong university of science and technology	stanford university	university of texas at austin
massachusetts institute of technology	nanyang technological university	rice university
microsoft	telefonica	university college london

SIGCOMM	SIGIR	SIGMOD
Affiliation_Name	Affiliation_Name	Affiliation_Name
google	yahoo	university of california berkeley
microsoft	florida international university	ibm
carnegie mellon university	microsoft	microsoft
stanford university	university of waterloo	google
telefonica	vienna university of technology	qatar airways
hp labs	university of amsterdam	university of southern california
universite catholique de louvain	wayne state university	duke university
university of wisconsin madison	istituto di scienza e tecnologie dell informazione	hong kong university of science and technology
university of massachusetts amherst	yandex	nanyang technological university
ibm	ibm	technische universitat munchen

Experiment 3 Result

Top 10 Field of Study with maximum number of Papers Selected in Various Conferences

KDD	MM	MOBICOM
Field_Name	Field_Name	Field_Name
Topic model	Software-defined networking	Wireless
Cluster analysis	Deep learning	Internationalization and localization
Social network	Crowdsourcing	Wi-Fi
Anomaly detection	Convolutional neural network	MIMO
Biological classification	Clos network	Wearable computer
Feature selection	Transport Layer Security	Channel state information
Collaborative filtering	Network management	Radio-frequency identification
Machine learning	Social media	Smartglasses
Crowdsourcing	Network congestion	Tracking
Information extraction	Remote direct memory access	Angle of arrival

SIGCOMM	SIGIR	SIGMOD
Field_Name	Field_Name	Field_Name
Software-defined networking	Evaluation	Fault tolerance
Clos network	Recommender system	Query optimization
Transport Layer Security	Information retrieval	Search engine indexing
Network management	Eye tracking	Multitenancy
Network congestion	Learning to rank	Data cleansing
Remote direct memory access	Collaborative filtering	Scalability
Bandwidth allocation	Sentiment analysis	Machine learning
Wireless	Personalization	Cloud computing
Hypertext Transfer Protocol over Secure Socket L...	Web search engine	Crowdsourcing
Load balancing	Hidden Markov model	Data mining

Experiment 4 Result

Top 10 Authors with maximum number of Papers Selected in Various Conferences

KDD	MM	MOBICOM
Author_Name	Author_Name	Author_Name
fei wang	quan guo	lili qiu
wei fan	rene kaiser	benjamin marlin
finale doshivelez	matteo varvello	mo li
jing jiang	anand raghuraman	he wang
kevin p murphy	bo li	souvik sen
nicholas d sidiropoulos	lazaros koromilas	you lizhao
james bailey	dinesh bharadia	david j perreault
yuqiang chen	qianqian xu	feng lu
ping zhang	charles clark	shyamnath gollakota
yan liu	chao zhang	stephanie gil

SIGCOMM	SIGIR	SIGMOD
Author_Name	Author_Name	Author_Name
matteo varvello	mihai lupu	michael j franklin
anand raghuraman	yexi jiang	amr ebaid
lazaros koromilas	ata turk	yatao li
dinesh bharadia	roy levin	patrick dantressangle
eiichi tanda	nemanja djuric	alexey reznichenko
charles clark	salvatore orlando	tim kraska
aditya akella	adam roegiest	marcelo arenas
michael kaminsky	maarten de rijke	yan liu
t telkamp	craig macdonald	ali ghodsi
kaveh razavi	andrew turpin	franz farber

Experiment 5 Result

Influence of Conference Location on Total no. of papers presented by Affiliations & selected

MM

Conf_Location	Total_No_of_Papers	Selected
Barcelona Spain	1038	**227**
Orlando / USA	1035	**237**
London	984	**274**
Singapore	201	
Portland USA - United States of America	60	
Hong Kong	12	
Sweden	3	

MOBICOM

Conf_Location	Total_No_of_Papers	Selected
Maui HI USA	703	**239**
Paris France	565	**243**
Miami Florida	417	**127**

Conclusion

Most Selected Affiliations

- Papers of Microsoft are mostly selected by conferences
- Other most selected Affiliations are Google, IBM, Massachusetts Institute of Technology, Tsinghua University, Carneige Mellon University etc

Most selected Field of Study by Conferences

- KDD conference focuses on Papers which are closely related to Machine Learning Fields like Classification, Clustering, Regression, Feature Identification, Data Mining etc
- MM & SIGCOMM conferences focuses on Papers which are closely relates to Network Field like Software Defined Networks, Neural Networks, Network Management, Social Media, Network Congestion etc
- MOBICOM conference focuses on Papers which are closely relates to Electronic Communication Field like WiFi, Wireless, Channel state, Wearable Computers, Radio Frequency, Tracking etc
- SIGIR conference focuses on Papers which are closely relates to Data Analysis and Artificial Intelligence Fields like Information Retrieval, Eye Tracking, Learning to rank, Collaborative filtering, Sentiment Analysis etc
- SIGMOD conference focuses on Papers which are closely relates to Business improvement oriented Fields like Data cleansing, Fault tolerance, Query optimization, cloud computing, scalability etc

Conference Location impact on Paper Participation

- Location impacted the MM conference heavily. Having location as Barcelona, Orlando and London encouraged more than 900 participation of papers from various Affiliations which dropped to less than 200 when location is changed to Singapore and further dropped to less than 65 when the location is changed to Portland, Hong-Kong and Sweden
- Location impacted the MOBICOM conference moderately. Paper participation is 500 to 700 when location of conference is Maui HI USA and Paris. But participation of papers dropped to less than 150 when location is changed to Miami Florida