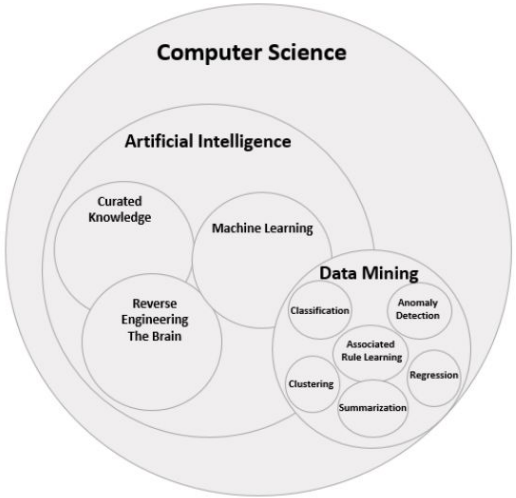
Comparison of Machine Learning Algorithms in Data Classification

Sarah Huang, Mridula Venkatasamy, Amy An, Julia Steed

Problem

- Today, extracting valuable information from raw data is essential in making effective business decisions.
- **Data Mining (DM)** and **Machine Learning (ML)** both play a vital role in extracting information from a huge amount of data
    - DM predicts outcomes or behaviors according to the future perspective and explores relationships.
    - ML is more effective to explore knowledge, validate the data and their behavior.



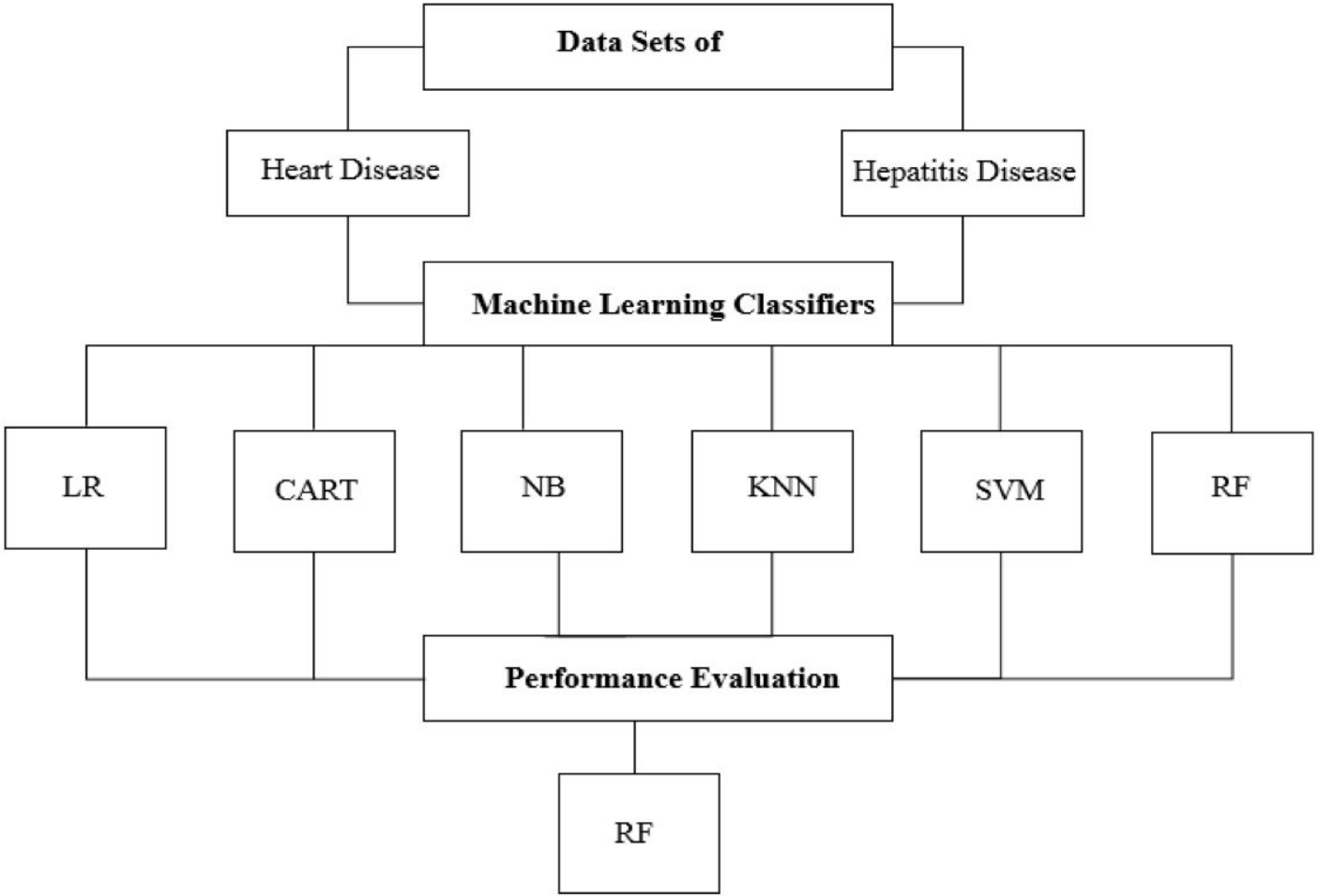
Which techniques more effectively analyze data, specifically health data?

Related Work

- Lots of research on DM classification methods and ML algorithms to predict diseases
  - DT, LL, NB, SVM, KNN, PCA, ICA classifier to analyze the kidney disease data
  - DT, SMO, BF Tree and IBK classifier to analyze breast cancer data
  - NB, J48 DT and Bagging algorithm, CART, ID3 (Iterative Dichotomized 3) and Decision Table, Logistics Classification, Multilayer Perceptron and SMO algorithms used to predict health of heart disease patients

TABLE I. CLASSIFIERS & DISEASE DISCUSSED IN LITERATURE REVIEW

References	Diseases	Classifiers	Year
[1]	Kidney	DT, NB, SVM,LL	2018
[2]	Kidney	KNN, NB, SVM,PCA, ICA	2017
[3]	Multiple	Map Reduce	2016
[4]	Breast Cancer	Classification(DT)	2014
[5]	Breast Cancer	SMO, IBK, BF	2007
[6]	Breast Cancer	Rep Tree, RBF,SL,DS	2017
[7]	Heart	J48DT, NB	2014
[8]	Heart	K-means, C4.5	2014
[9]	Heart	CART, ID3, Decision Table	2013
[10]	Heart	Weka Tool	2013
[11]	Kidney	DT, ANN, LR	2014
[12]	Kidney	ANN	2013
[13]	Heart & Kidney	Apriori,K-means	2014


Data Sets of
Heart Disease
Hepatitis Disease
Machine Learning Classifiers
LR
CART
NB
KNN
SVM
RF
Performance Evaluation
RF

Methodology - Evaluation parameter

- In this paper, they review 4 important evaluation parameters in data mining: **Accuracy, precision, recall, and F measure.**
- Accuracy is the number of accurately classifies instances divided by a total number of instances in the dataset.
- Precision is the average probability of relevant retrieval.
- Recall is the average probability of complete retrieval.
- F-measure is calculated using precision and recall.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \dots (1)$$

$$Precision = \frac{TP}{TP + FP} \dots (2)$$

$$Recall = \frac{TP}{TP + FN} \dots (3)$$

$$F\ Measure = \frac{2*(Precision*Recall)}{Precision*Recall} \dots (4)$$

Methodology - Dataset

- The researchers used datasets of heart disease and hepatitis.
- The heart disease dataset contained 14 attributes and 303 instances.
- The hepatitis dataset contained 20 attributes and 155 instances.
- Taken from UCL repository.

Methodology - Logistics regression

- LR is the predictive analysis to conduct on discrete (binary) values based on a specified set of independent variables.
- Describes the data and clarifies the relationship between one dependent variable and independent variables
- Predicts occurrence probability by fitting data logit function.

$$P = \frac{e^{(b0+b1*x)}}{1 + e^{(b0+b1*x)}} \dots (5)$$

Methodology - Decision tree

- DT is a decision supporting tool that is represented by a tree-like graph or model of decision that represent the possible outcomes
- Helps to identify a strategy and can be particularly useful for decision analysis to reach the goal more effectively
- The data is split into 2 homogeneous set, then decision analysis is performed.

Methodology: Machine Learning Classifiers
Naive bayes

- Probabilistic classifiers based on Bayesian theorem with naive independence assumption between the predictors or features
- Assumes that the existence of particular feature is not related to existence of any other feature in class
    - All features to contribute independently to probability identification

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Methodology: Machine Learning Classifiers
K-Nearest Neighbours (KNN)

- Most commonly used in classification problems
  - Used for regression as well
- Stores all existing cases then classified new cases by the majority votes of its neighbors, measured by distance function
  - Euclidean
  - Manhattan
  - Minkowski

Methodology: Machine Learning Classifiers
Support Vectors Machine (SVM)

- Classification and regression algorithms
- Every data item is plotted in n-dimensional space
  - Number of dimensions = number of features or attributes
- Drawing a line or finding optimal hyperplane that separate two classes after plotting all data

Methodology: Machine Learning Classifiers
Random Forests (RF)

- Learning technique for regression, classification, and more
- Make a multitude of decision trees at training stint and output the mean prediction (regression) or mode of classes (classification) of the distinct trees
- In the paper, forest indicates classification of the class that has most “votes”
- For every tree,
    - Sample of N objects in training set randomly taken
    - If input variable M is bigger than the randomly selected node, split
    - No trimming used

Experimental Results

- Evaluated classifiers in terms of accuracy, precision, recall, and F measure
- Random Forest (RF) outperformed the other 5 classifiers for both heart and hepatitis disease prediction
- RF achieved 83% accuracy for heart disease, followed by SVM (82% accuracy)
- RF achieved 85% accuracy for hepatitis, followed by Linear Regression (82% accuracy)
- Naive Bayes performed the worst for both diseases

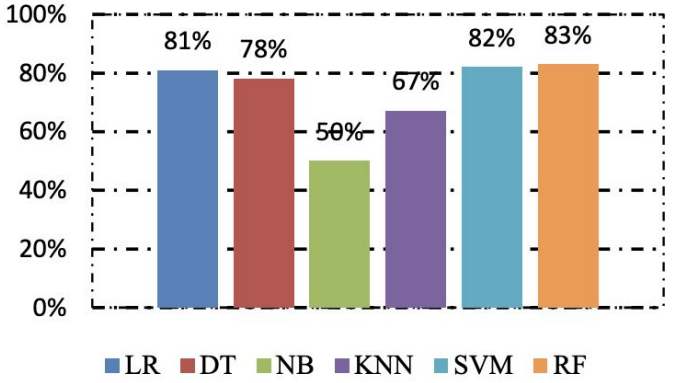


Figure 5. Heart Disease Accuracy

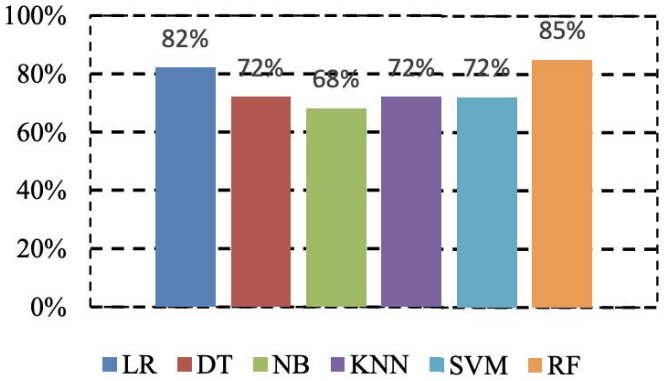


Figure 9. Hepatitis Disease Accuracy

Conclusion

- Machine learning and data mining are applicable to many fields, such as healthcare, and can be highly effective at making predictions
- Different ML classifiers can vary greatly in performance on the same dataset
- Random Forest consistently performed better than the other classifiers on both datasets

References

- C. A. Ul Hassan, M. S. Khan and M. A. Shah, "Comparison of Machine Learning Algorithms in Data classification," 2018 24th International Conference on Automation and Computing (ICAC), Newcastle Upon Tyne, UK, 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748995. keywords: {Diseases;Kidney;Heart;Data mining;Support vector machines;Machine learning;Classification algorithms;Data Mining;Machine Learning;Data Analysis},