Data Science and Analytics

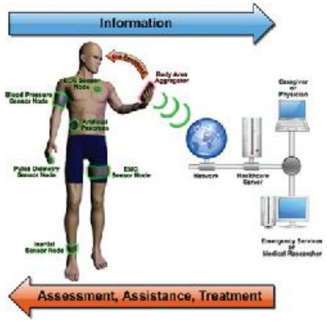
Dr. Manas Ranjan Patra
Professor of Computer Science
Berhampur University, Berhampur, India

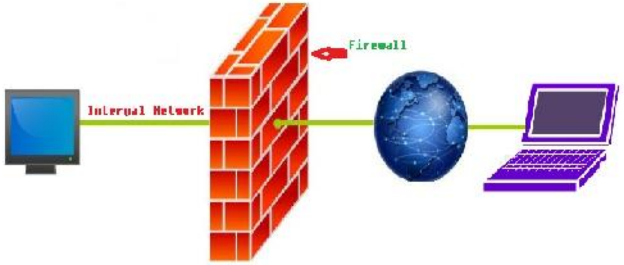
DRIP Syndrome ....

We are Data Rich …….

But Information (Insight) Poor….

Data Sources





- Transaction Databases → Bank, Shopping Mall, PoS
- Social Media Data → Product Reviews, Facebook, Tweets, LinkedIn
- Wireless Sensor Data → Real-time Monitoring, Internet of Things
- Software Log Data → system log, network monitoring (Firewall/IDS), Cookies

“Data is the New Oil”

– World Economic Forum 2011 Report

“Data is just like crude oil. It’s valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value.”

Turing award winner Jim Gray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven)

What can you do with the data?

- Build organizational knowledge
- Build data models to understand the relations (explicit/implicit)
- Predict
  - Stock market fluctuations
  - Traffic dynamics
  - Weather forecasting
  - Customer churn (switching from one company to another)
  - Who will form the next govt.?
- Strategic decision-making

5 Vs of Data

- Volume (size/amount)
- Velocity (speed of generation, change over time)
- Variety (diversity in data types, format, source)
- Veracity (Data Quality)
- Value (Information for Decision Making)

Emergence of Data Science

- **Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms (both structured and unstructured) - **Wikipedia**
- Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

Nature of data processing depends on the purpose!

- Association
- Outlier detection
- Understanding data patterns
- Classification
- Clustering
- Building Decision Trees
- Rule mining
- Machine learning

The ability to automatically learn from data & build models

Word of Caution :

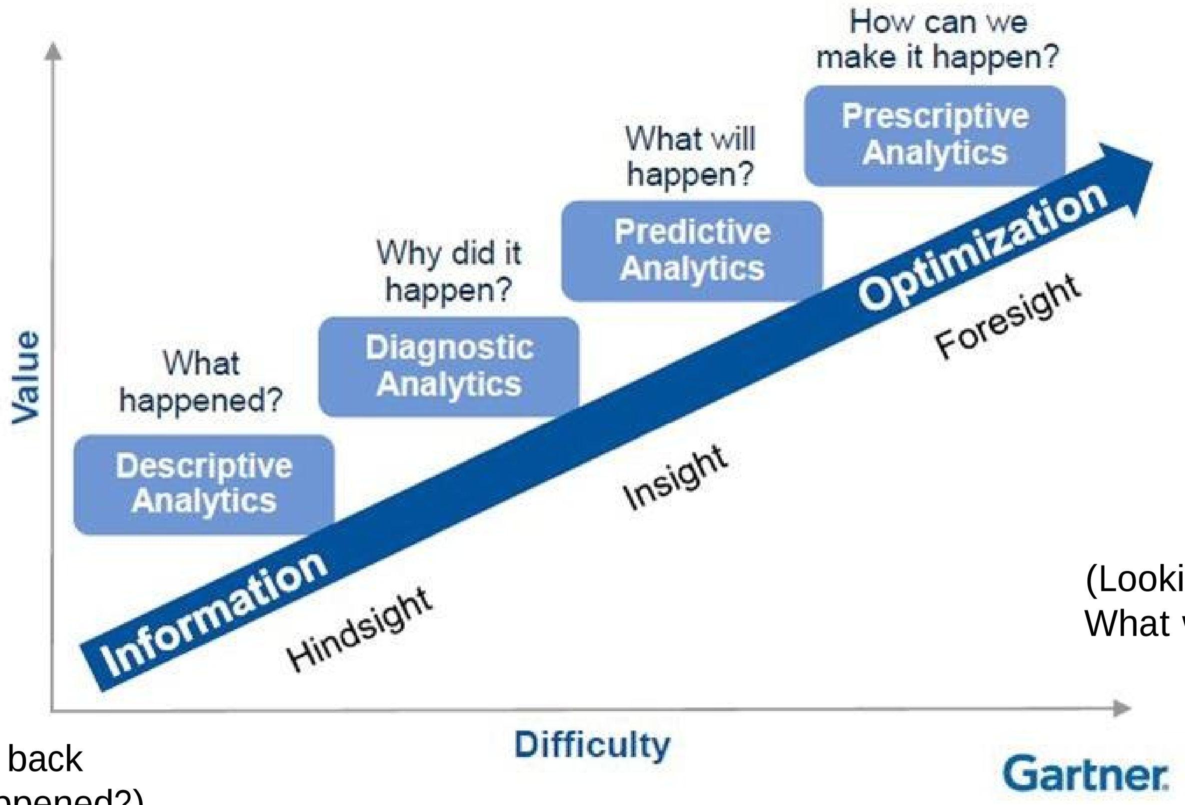
Just don’t jump on a technique !!

What is Data Analytics?

Data Analytics :

- “is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making”. - Wikipedia
- "leverage data in a particular functional process (or application) to enable context-specific insight that is actionable.“ – Gartner

Data Analytic Capabilities



(Looking forward
What will happen?)

(Looking back
What happened?)

“The greatest enemy of knowledge is not ignorance, it is the illusion of knowledge.”

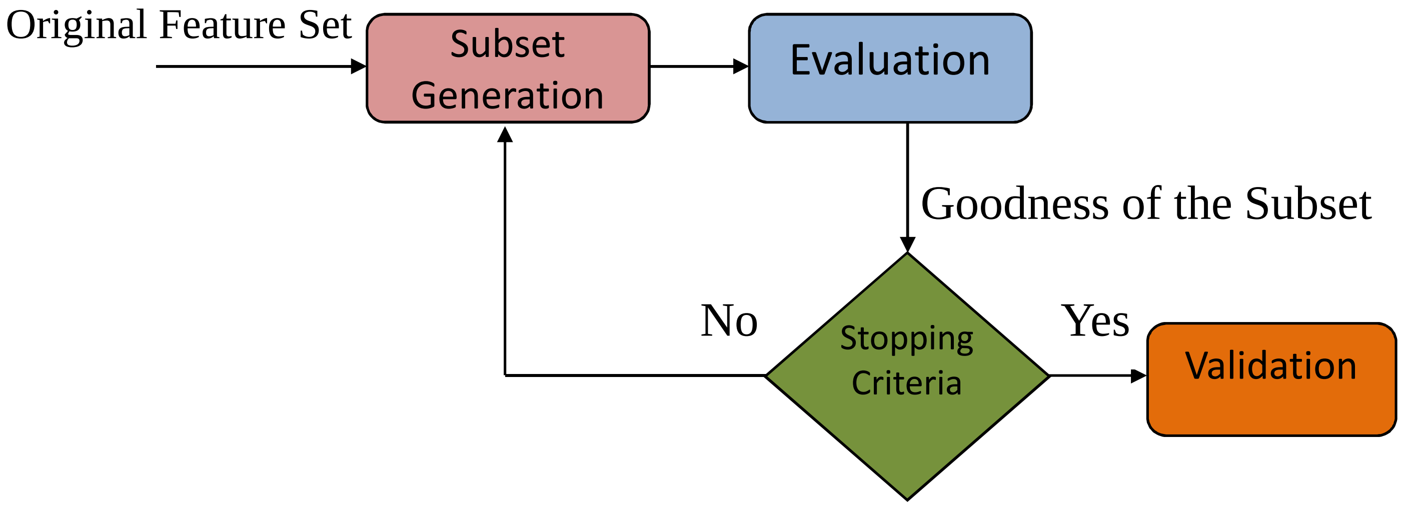
-Stephen Hawking

Data Cleansing

- Filling in missing data
- Detecting and removing outliers
- Smoothing (removing noise by averaging values together)
- Filtering, sampling (keeping only selected representative values)
- Feature extraction

Feature Selection Techniques

- Not all features (attributes) are relevant to the intended processing.
- Therefore, find the most relevant subset of attributes.



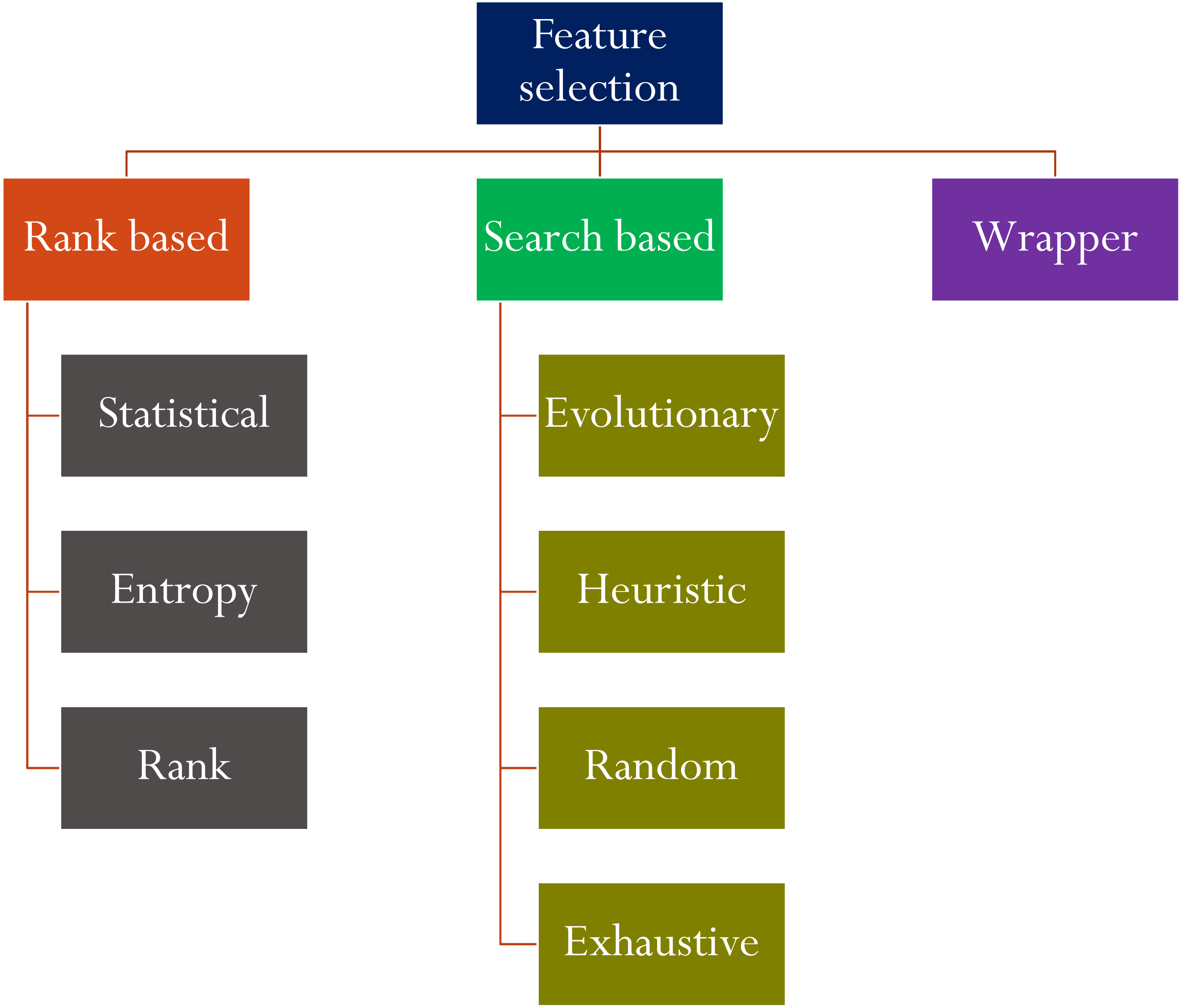
General Feature Selection Process

Benefits of Feature Selection

- Reduces the size of the problem.
- Reduces the requirement of computer storage.
- Reduces the computation time.
- Reduction in features can improve the accuracy as irrelevant attributes are removed.

Approaches for Feature Selection

1. **Rank based feature selection** (Filtering approach)
   - The variables are assigned with a score using suitable ranking criterion and those having score below a threshold are dropped.
   - Computationally cheaper but ignores dependencies among the features. Thus, the selected subset might not be optimal.
2. **Search based feature selection** (Embedded approach)
   - Searches for an optimal subset of features
   - Less computationally intensive than wrapper method
3. **Wrapper Method**
   - Evaluates attribute sets by using a learning scheme. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes.
   - Considers feature dependencies but slow


Feature selection
Rank based
Search based
Wrapper
Statistical
Entropy
Rank
Evolutionary
Heuristic
Random
Exhaustive

Statistical Based

- Relief-F
- One-R
- Chi-Squared

Entropy based Methods

Entropy is commonly used in information theory which characterizes the purity of an arbitrary collection of examples. The entropy is considered as a measure of system's unpredictability.

- Information Gain
- Gain Ratio
- Symmetrical Uncertainty

Rank Method

- Principal Component Analysis (PCA)(Feature Extraction)

Evolutionary Search Method (Nature Inspired Feature Selection)

- Ant Search (AS)
- Genetic Search (GS)
- Particle Swarm Optimization (PSO) Search

Heuristic / Informed Search

- Best First Search
- Greedy Stepwise
- Linear Forward Selection (Extension of Best First)
- Hill Climber
- Combined Hill Climber

Random Search

- Feature Subset Harmony Search
- Feature Vote Harmony Search

Exhaustive Search

- Performs an exhaustive search through the space of attribute subsets starting from an empty set of attributes. Reports the best subset found.

Wrapper Method

- Wrapper Subset Evaluator

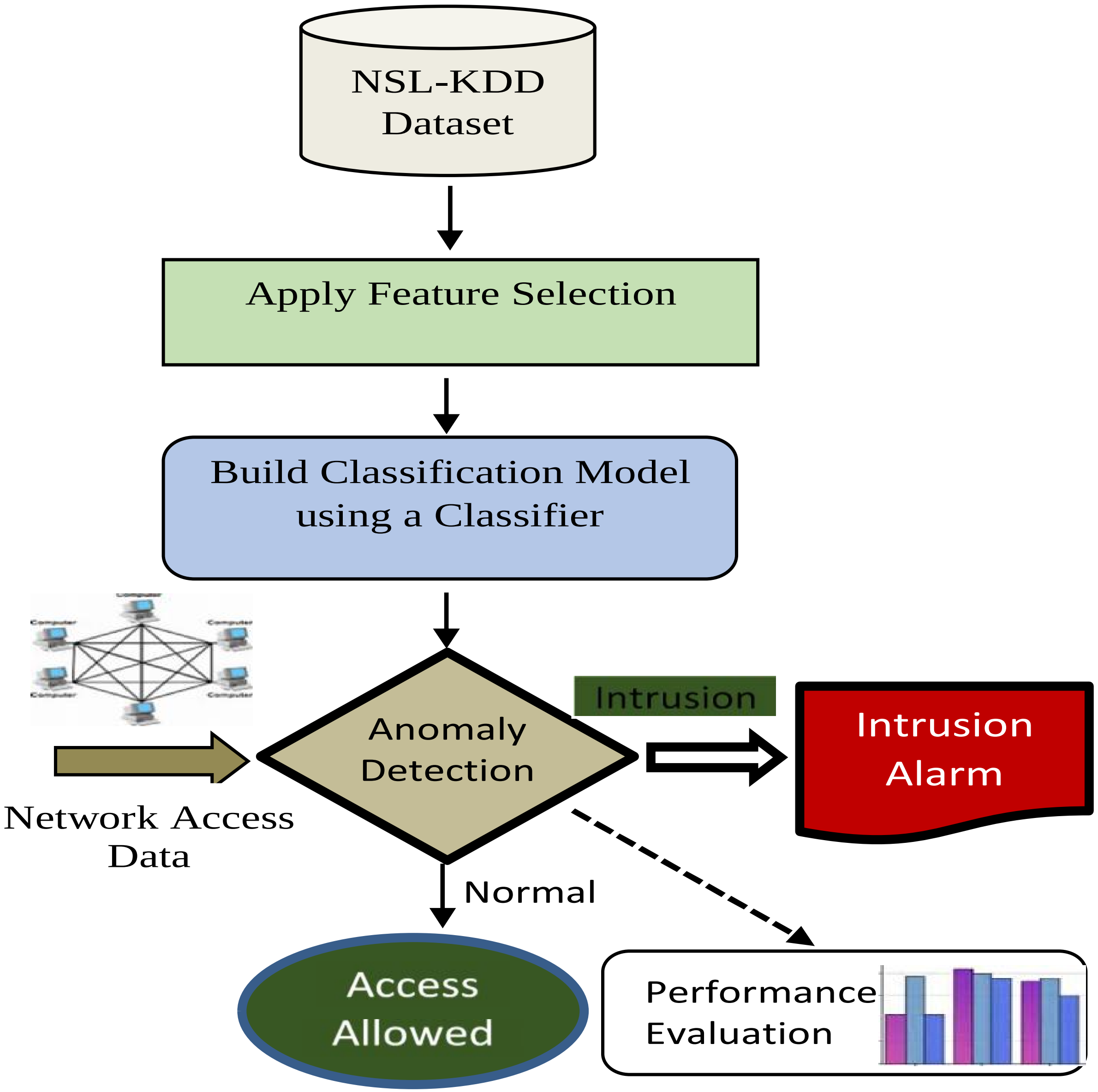
- The wrapper approach conducts a search in the space of possible parameters. The search requires a state space, an initial state, a termination condition and a search engine.
- The goal of the search is to find the state with the highest evaluation using a heuristic function.
- The accuracy of the learning scheme in finding a subset of most relevant attributes is evaluated using Cross Validation technique.

Cross Validation Technique

- Cross validation calculates the accuracy of the model by separating the data into two different populations, a training set and a testing set.
- In n-fold cross-validation the dataset is randomly partitioned into n mutually exclusive folds
- E.g., in 10-fold cross validation, a given dataset is partitioned into 10 subsets (9 subsets are used for training and the 10th subset for testing). This cross-validation process is then repeated 10 times (the number of folds).

Experiments to Study

Impact of Feature Selection


NSL-KDD Dataset
Apply Feature Selection
Build Classification Model using a Classifier
Anomaly Detection
Network Access Data
Intrusion
Intrusion Alarm
Normal
Access Allowed
Performance Evaluation

Impact of Feature Selection

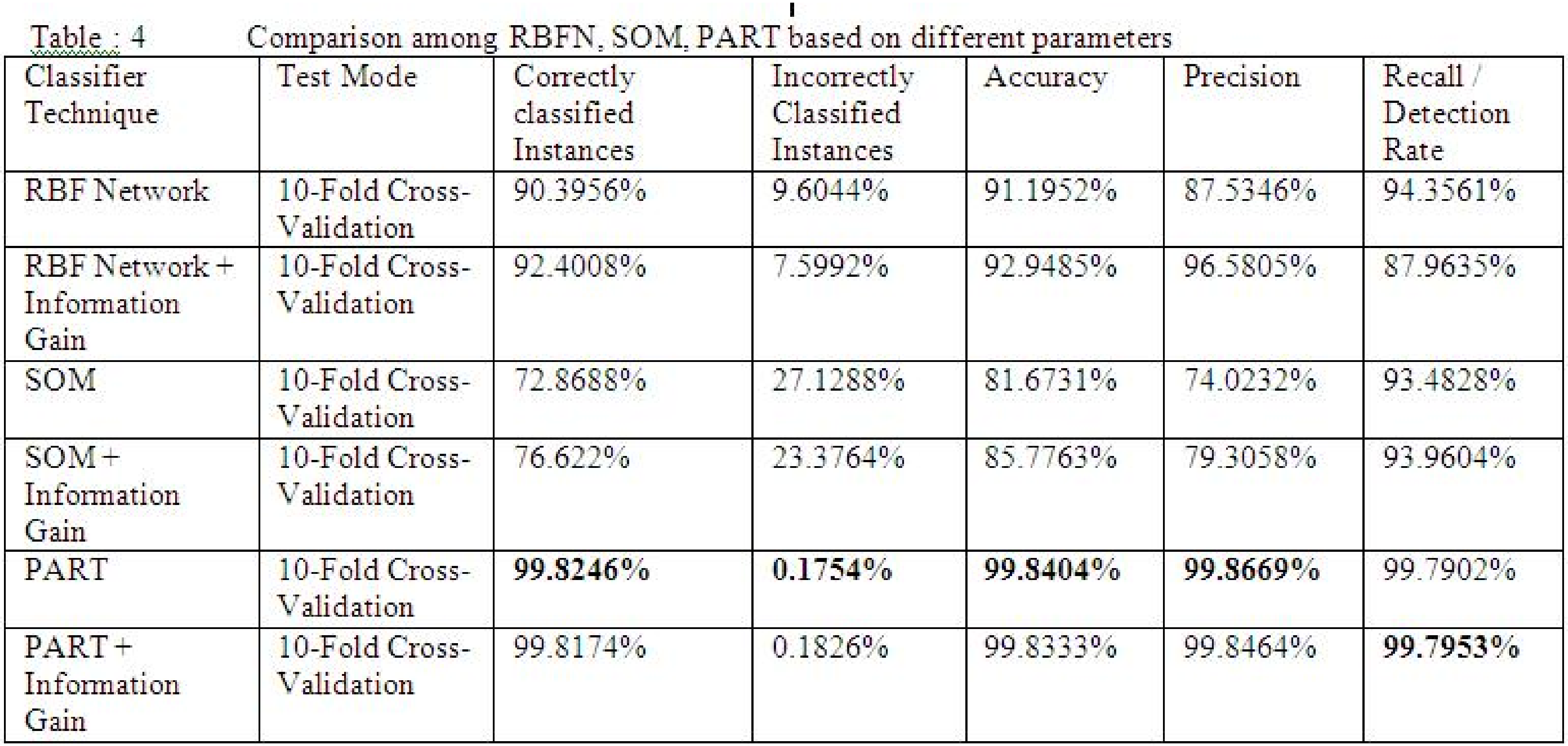


Table : 4 Comparison among RBFN, SOM, PART based on different parameters

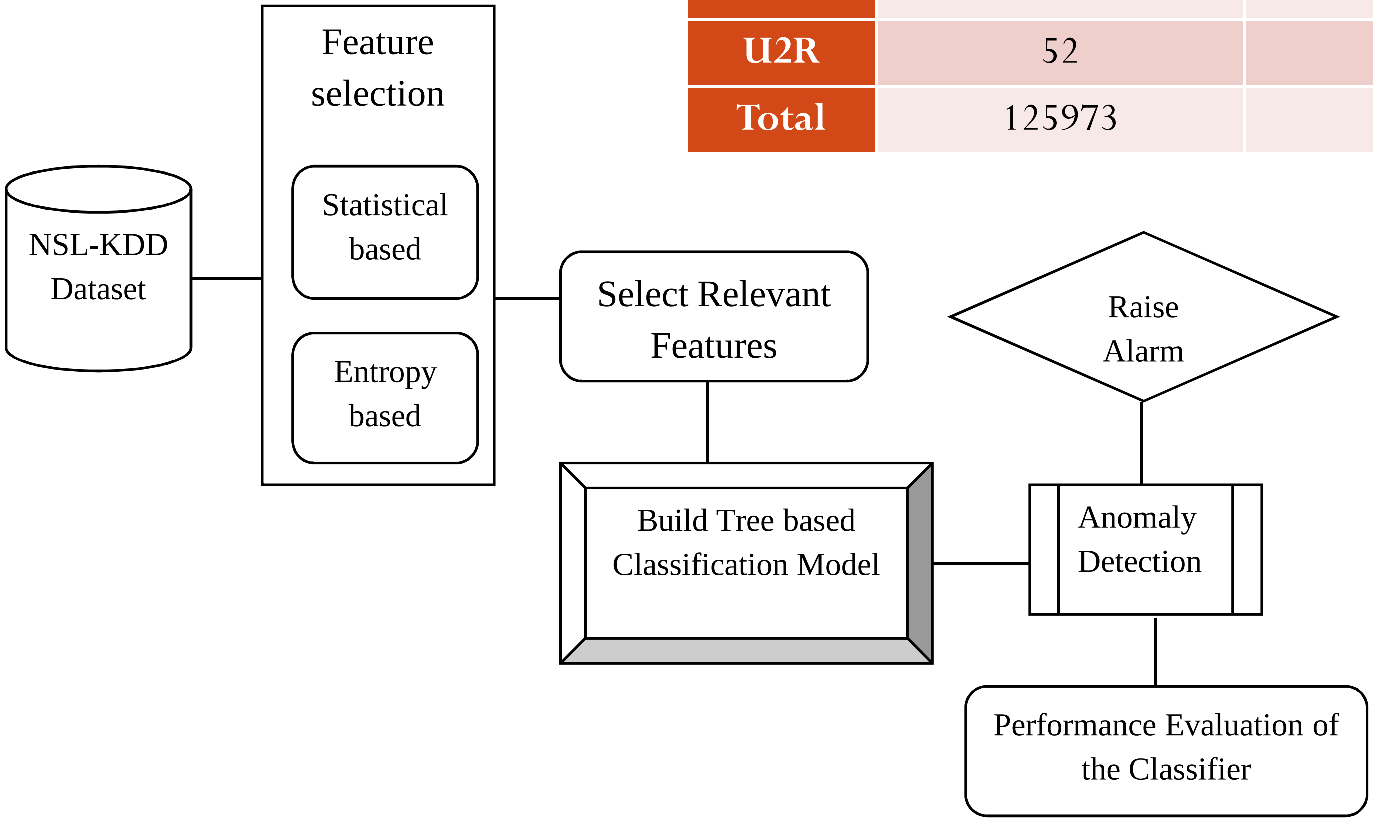
Classifier Technique	Test Mode	Correctly classified Instances	Incorrectly Classified Instances	Accuracy	Precision	Recall / Detection Rate
RBF Network	10-Fold Cross-Validation	90.3956%	9.6044%	91.1952%	87.5346%	94.3561%
RBF Network + Information Gain	10-Fold Cross-Validation	92.4008%	7.5992%	92.9485%	96.5805%	87.9635%
SOM	10-Fold Cross-Validation	72.8688%	27.1288%	81.6731%	74.0232%	93.4828%
SOM + Information Gain	10-Fold Cross-Validation	76.622%	23.3764%	85.7763%	79.3058%	93.9604%
PART	10-Fold Cross-Validation	**99.8246%**	**0.1754%**	**99.8404%**	**99.8669%**	99.7902%
PART + Information Gain	10-Fold Cross-Validation	99.8174%	0.1826%	99.8333%	99.8464%	**99.7953%**

NSL-KDD Data set with 41 feature attributes (38 numeric and 3 symbolic).

Total number of records 1,25,973 (67,343 normal & 58,630 attacks)

Data distribution

Class	Number of Records	% of occurrence
Normal	67343	53.48%
DOS	45927	36.45%
Probes	11656	9.25%
R2L	995	0.78%
U2R	52	0.04%
Total	125973	



NSL-KDD Dataset

Feature selection

Statistical based

Entropy based

Select Relevant Features

Build Tree based Classification Model

Anomaly Detection

Raise Alarm

Performance Evaluation of the Classifier

Confusion Matrix

		Predicated Class	
		Normal	Attack
Actual Class	Normal	True Negative (TN)	False Positive (FP)
	Attack	False Negative (FN)	True Positive (TP)

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TN+TP+FN+FP}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Detection Rate or Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

$$\text{False Alarm Rate} = \frac{\text{FP}}{\text{TN+FP}}$$

Experiment -2: Artificial Neural Network based Classification Techniques

- *Self-Organizing Map (SOM)*
- *Projective Adaptive Resonance Theory (PART)*
- *Radial Basis Function Network (RBFN)*
- *Sequential Minimal Optimization (SMO)*

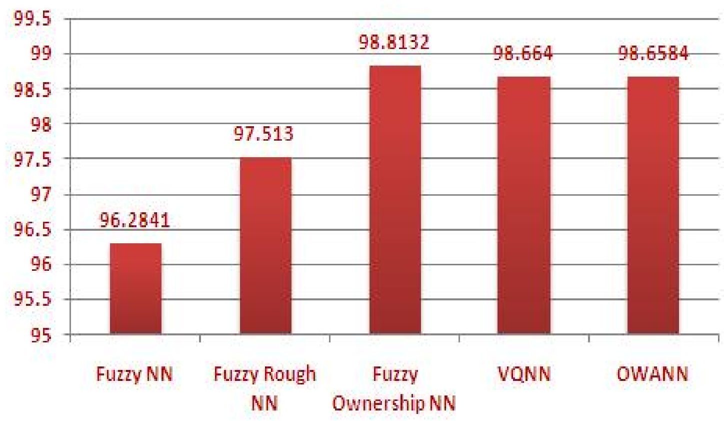
Feature Selection Method	**Classifier Techniques**	Accuracy in %	Precision in %	Detection Rate / Recall in %	False Alarm Rate in %
Ant Search	SOM	85.8446	78.2802	96.3141	23.267
	PART	**99.5475**	**99.8681**	**99.4696**	**0.3757**
	RBFN	91.0314	93.9921	86.2425	4.7993
	SMO	91.8832	96.6266	85.5466	2.6001
Random Search	SOM	84.7166	77.9295	93.7029	23.1048
	PART	**99.8381**	**99.8413**	**99.8107**	**0.1381**
	RBFN	94.094	95.4941	91.634	3.7643
	SMO	96.6453	96.9388	95.8178	2.6343

Experiment -3: Fuzzy Rough set based Classification Techniques with Wrapper subset evaluator feature selection

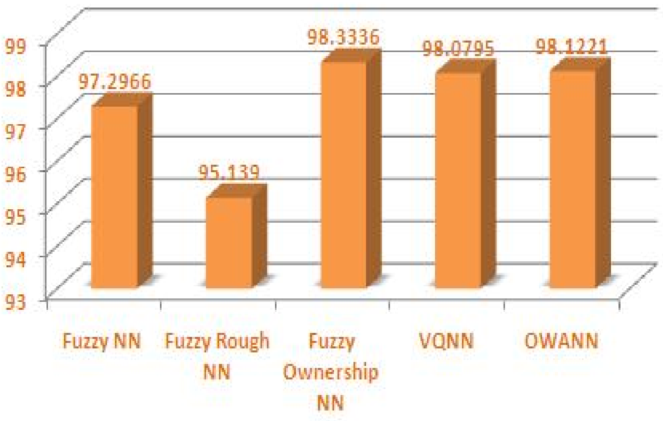
- Fuzzy Nearest Neighbour
- *Fuzzy-Rough Nearest Neighbour*
- *Fuzzy-Rough Ownership NN*
- *Vaguely Quantified Nearest Neighbour*
- *OrderedWeighted Average Nearest Neighbour*

Classifier Techniques	Evaluation Criteria			
	Accuracy in %	Recall or Detection Rate in %	Precision in %	False Alarm Rate in %
Fuzzy NN	96.2841	97.2966	94.8521	4.5973
Fuzzy Rough NN	97.513	95.139	**99.4952**	**0.4202**
Fuzzy Ownership NN	**98.8132**	98.3336	99.1283	0.7529
VQNN	98.664	98.0795	99.0407	0.8271
OWANN	98.6584	98.1221	98.6584	0.8746

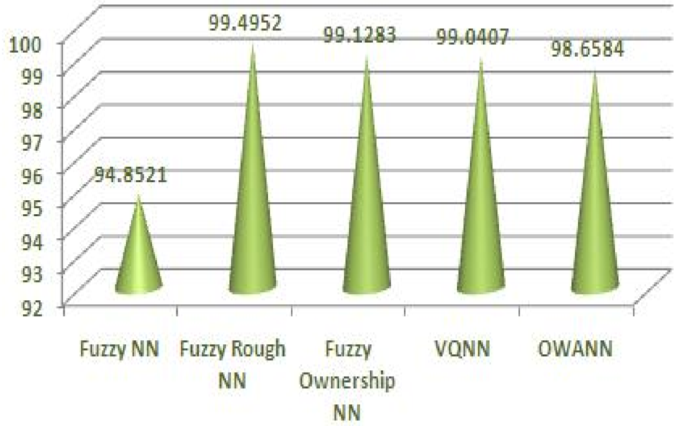
Accuracy in %



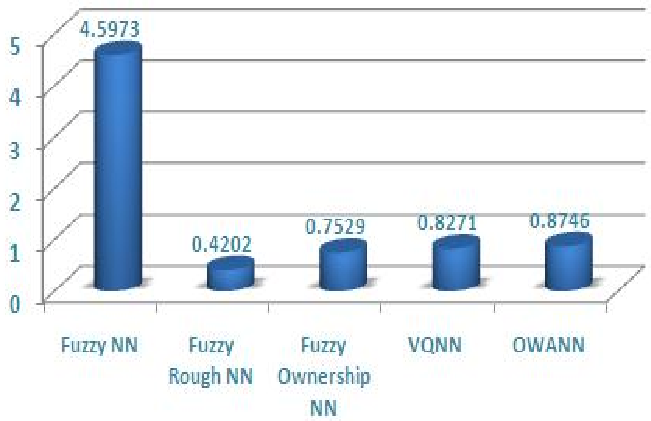
Detection Rate in %



Precision in %



False Alarm Rate in %



Rule based Classification

- Ripple Down Rule Learner (RIDOR)
- Non-Nested Generalized Exemplars (NNGE)
- JRip
- Decision Table/Naïve Bayes (DTNB) Classifier

Supervised Learning (i.e. with a "teacher")

- Radial Basis Function Network (RBFN)
- Back-Propagation Algorithm
- Hopfield Network
- Support Vector Machine (SVM)
- Naïve Bayes Classifiers or Bayesian classification
- Decision Tree-Based Algorithms (J48, NB tree, Random forest , Random Tree, REP tree, Simple CART, Best First decision Tree, Function Tree)
- K-Nearest Neighbor Algorithm (Distance-based algorithm)

Contd..

- Genetic Algorithm (GA) (Evolutionary computing methods and are optimization –type algorithms)
- Apriori Algorithm for Association Rule Learning
- Fuzzy Rough set Nearest Neighbour Algorithms (Fuzzy Nearest Neighbour (FNN), Fuzzy-Rough Nearest Neighbour (FRNN), Fuzzy-Rough Ownership Nearest Neighbour (FRONN), Vaguely Quantified Nearest Neighbour (VQNN), and Ordered Weighted Average Nearest Neighbour (OWANN).

Contd..

- Artificial Immune Recognition System (AIRS1, AIRS2, clonalg, Clonal Selection Classification Algorithm (CSCA)) (Evolutionary computing method)
- Rule Learning techniques (Conjunctive rule, decision table, Decision Table / Naïve Bayes (DTNB), JRip, NNGE, RIDOR)

Rule Learning Techniques

- Conjunctive Rule
- Decision Table
- DTNB (Decision Table / Naïve Bayes)
- Jrip
- NNGE
- RIDOR

Tree based Techniques

- J48
- NB Tree
- Random Forest
- Random Tree
- REP Tree (Reduced-Error Pruning)
- Simple Cart
- Best First Decision Tree (BF Tree)
- Function Tree

Unsupervised Learning (i.e. without a “teacher”)

- Kohonen Self-Organizing Map (SOM) / Topology Preserving Maps
- Adaptive Resonance Theory (ART)
- K-Means (Partition algorithms)
- Hierarchical Algorithms (Agglomerative and Divisive algorithms)
- Expectation-Maximization Algorithm (EM)
- Learning Vector Quantization (LVQ)
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) – useful for Clustering large database

Hybrid Systems

- Neuro-genetic system
- Fuzzy-neural system
- Fuzzy-genetic system
- Neuro-fuzzy hybrid system

"If we have data, let's look at data. If all we have are opinions, let's go with mine."

- Jim Barkesdale, CEO of Netscape

Q & A

Thank You!