ANKAM

Machine Learning for Drug Design

Big Data Meets Medicine

OCTOBER, 2015

Outline

- Executive Summary
- Problems in Drug Design and Solutions Offered
- Introduction to Machine Learning
- Dimensionality Reduction and Feature Extraction
- Decision Trees and Random Forests
- Support Vector Machines
- Neural Networks
- Deep Learning and Convolutional Neural Networks (CNN)
- Merck Challenge and Virtual Screening
- About Us
- Reference

Executive Summary

- ANKAM applies Statistical Learning and Computational Models to problems in Finance.
- We are now looking to apply Machine Learning in the field of Healthcare with current focus on Personalized Medicine.
- In this document, we will
  - describe some cutting edge Machine Learning techniques
  - show how we use these techniques in Finance
  - illustrate how they can be useful for analysis of Big Data in the Drug Discovery process

“Pain Points” in the Drug Design Process

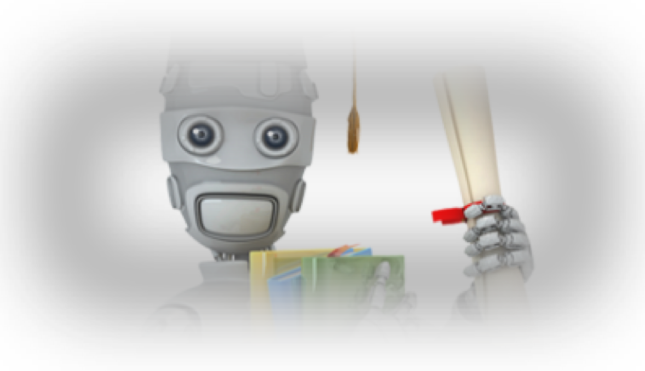
- Drug development process is long and costly with low probability of success.
- Cost of developing a drug is between 2 and 3 billion dollars
- The entire process can take up to 10 years.
- Many Clinical trials fail but this does not necessarily mean the drug was bad.
- Analysis of the Clinical Trials data may lead us to golden nuggets. For example, FDA failed drug candidates could be re-analyzed to determine if there is a subpopulation of patients for which the drug-candidate is effective.
- For every patient that responds to treatment by a drug, between 4 to 24 patients do not respond to the same drug- leading to “trial and error” approach in treatment.

Addressing the “Pain Points” in Drug Design

- Next Generation Sequencing (NGS) is becoming more mainstream and producing an avalanche of “Big Data”.
- The data are not only big in size but also have very high dimension. Traditional Statistical techniques are ill suited to handle this type and volume of Data.
- Predictive Algorithms and Techniques from Machine Learning which have proved successful on Image Recognition and Natural Language Processing can help in identifying important biomarkers, which can lead to targeted therapies.
- Intersection of Machine Learning with Molecular Biology to analyze Big Data is the key to Personalized Medicine.

Background : Introduction to Machine Learning

- Machine Learning is
  - What Gmail uses to identify spams
  - What Amazon uses to automatically give you product recommendations
  - What Facebook uses to identify faces in photos
  - What LinkedIn uses to recommend connections
  - What Google Translate uses to deliver instant translations
  - What Netflix uses to recommend Movies
  - What powers Apple's Siri virtual personal assistantance system
  - ...



- In layman's term, it is a class of algorithms that adjust and learn from data to take actions in the future.

Background : Applications of Machine Learning

Machine Learning can solve a wide array of real-life problems:

Finance

- Financial modeling
- News Sentiment Analysis
- Trading signals
- Fraud detection

Biology

- Patient behaviors and disease
- New Drug Design
- Personalized Medicine
- computer-aided diagnosis

Retail and Marketing

- Item recommendation
- Market segmentation
- Targeted Ads
- Buyer Sentiment Analysis

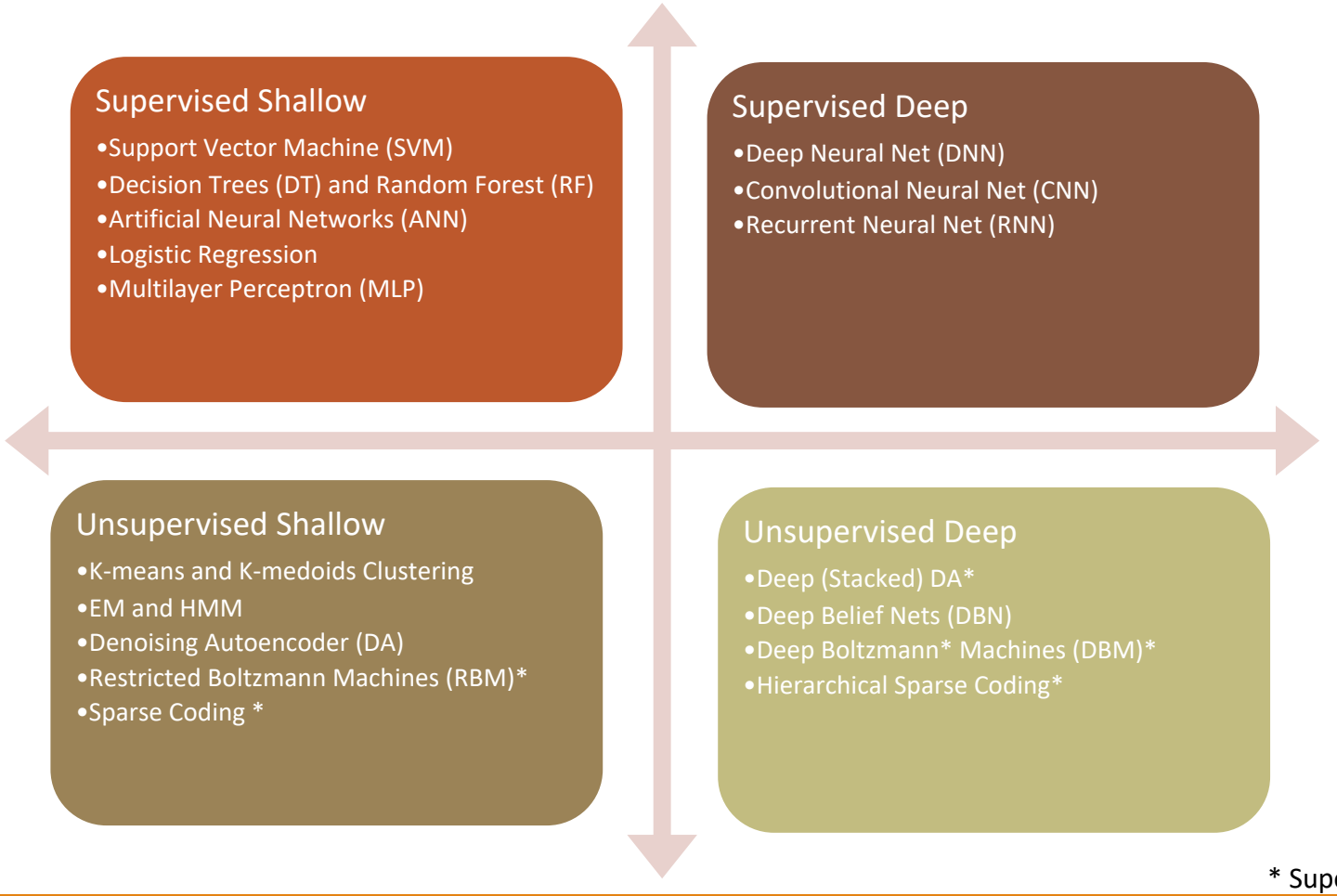
Vision

- Face recognition
- Handwriting Recognition
- Image Segmentation
- General Object Classification
- Personal Assistant

Language

- Language Translation
- Text Classification
- Sentiment Analysis
- Question Answering
- Summarization

Taxonomy of Popular Machine Learning Algorithms



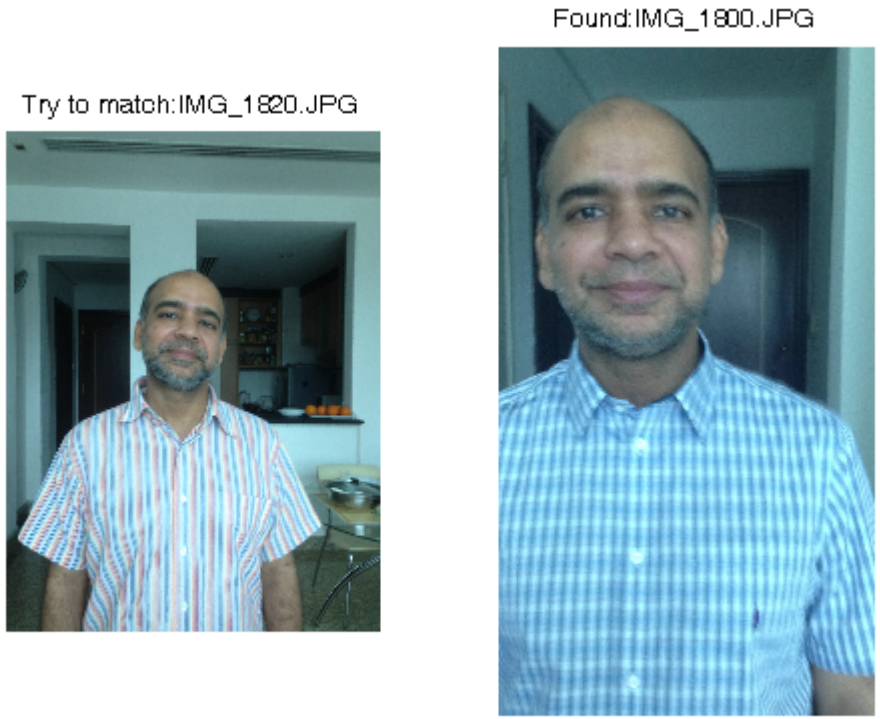
* Supervised version also exists

Feature Extraction and Dimensionality Reduction

- Dimensionality reduction is an important pre-processing step in various machine learning techniques, which transforms the data from high dimensional space into a lower dimensional feature space. This step aims at removing redundant and irrelevant features, which results in improved training efficiency, result comprehensibility and prediction accuracy.
- Commonly used Dimensionality Reduction techniques include but not limited to
  - Principal Component Analysis (PCA) and Kernel PCA
  - Linear Discriminant Analysis (LDA)
  - Manifold Learning (Isomap, locally linear embedding (LLE), Hessian LLE, etc.)

How ANKAM uses PCA for Human Face Recognition

Figure 3. Human Face Recognition using Eigenfaces



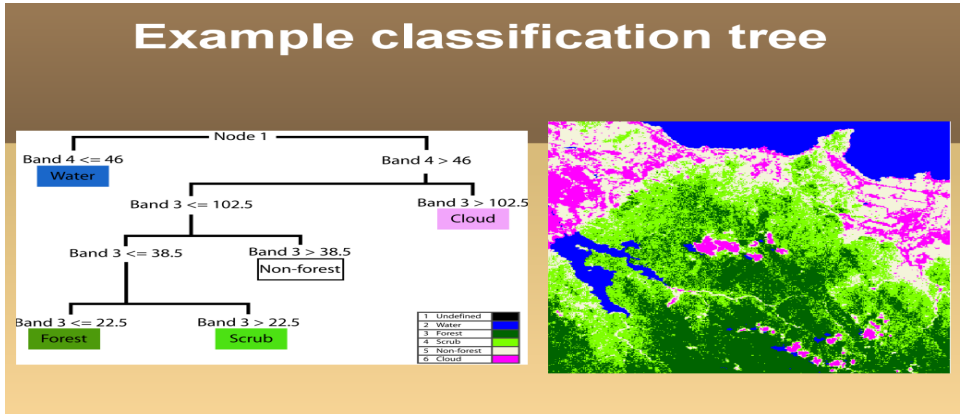
Eigenfaces Algorithm

We performed the face recognition using Eigenfaces (PCA) approach. First of all, we train the model using a training set of photos, which contain the person to be recognized. We apply PCA to this matrix to reduce the dimension of each face to K eigenfaces. In the recognition step, we project the photo into the eigenspace, and match the photo in the database with the smallest distance measure.

This algorithm is able to match the two faces even though the shirt, background, tilt of the face are different.

What is a Decision Tree?

- Decision tree is a predictive model that uses a set of binary rules applied to calculate a target value. It can be used for classification or regression.
  - In Classification, categorical variables are used. E.g., whether a tumor is benign or malignant.
  - In Regression applications, continuous variables are used. E.g., the molecular activity for a compound on a target.
- A recursive tree generating algorithm is used to learn a Decision Tree, and the "best" branching policy is the one that results in the largest information gain. To avoid overfitting, pruning is used. Techniques like cross validation can be used for model validation and testing.
- Advantage of decision tree
  - Easy to interpret
  - Robust as regard to outliers in the training data
  - Prediction is fast once rule is developed

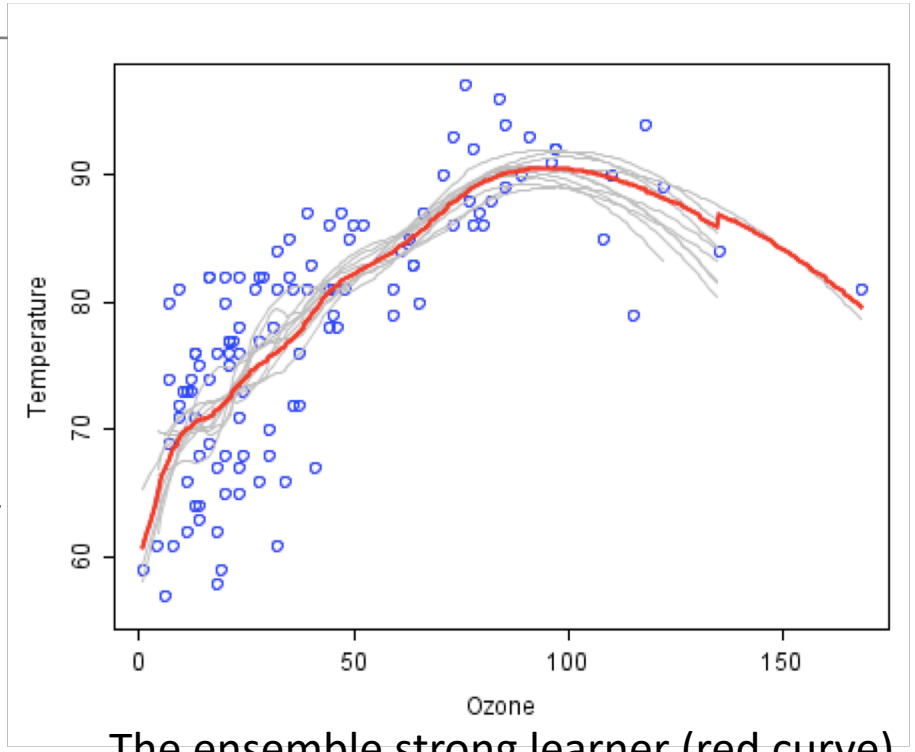


What is a Random Forest (RF)?

- The **Random Forest** (Breiman, 2001) [1] is an Ensemble approach that uses many Decision Trees (weak learners) to form a refined final classification or regression (strong learner) that is more stable and accurate than all the individual decision trees.

- To train a RF model, a different subset of the training samples is selected (approx. 2/3 of the original data) with replacement to train each tree. Final classification is made by majority voting and regression is made by tree averaging.

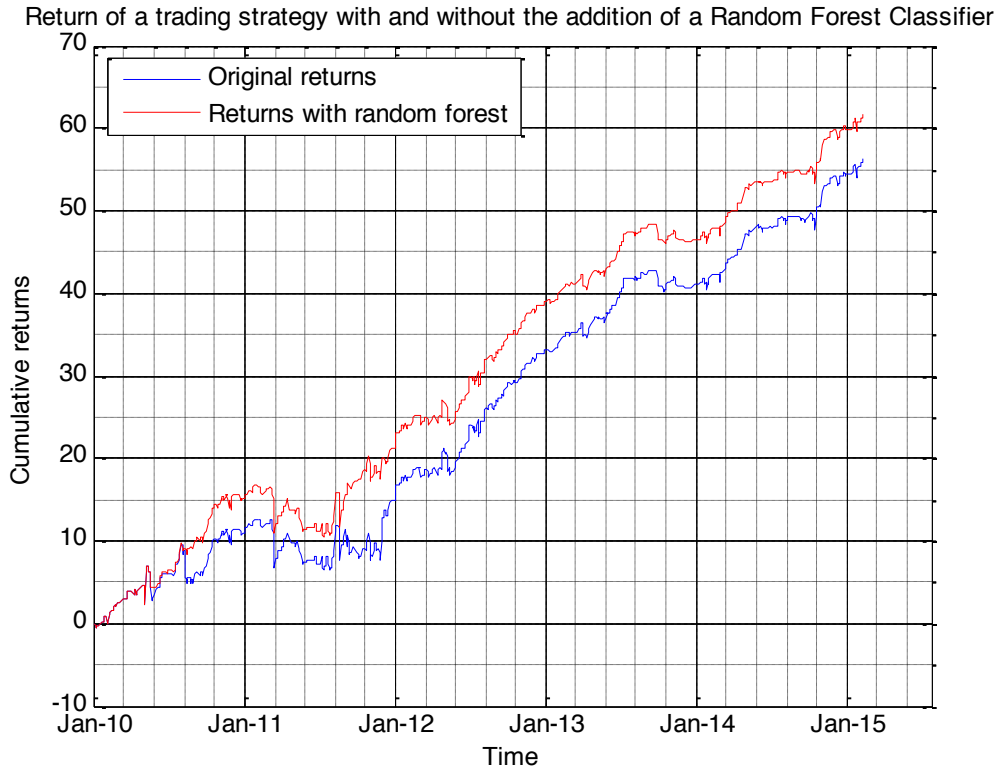
- In contrast to a single Decision Tree, RF is robust to data overfitting and thus no tree pruning is required. It is also robust as regard to outliers in the training data.



The ensemble strong learner (red curve) trained using Random Forest.

How ANKAM uses Random Forest to improve Models for Finance

Figure 1. Trading Signals enhanced by Random Forest



ANKAM RF Classifier

We form a trading algorithm that continuously makes trading decisions. The algorithm has no knowledge of a specific set of variables ("withheld variables"). We then train a random forest classification model using the withheld variables. Based on the output of the trained model, we accept or reject certain trading signals. This random forest filter is able to improve the original model substantially.

What are Support Vector Machines?

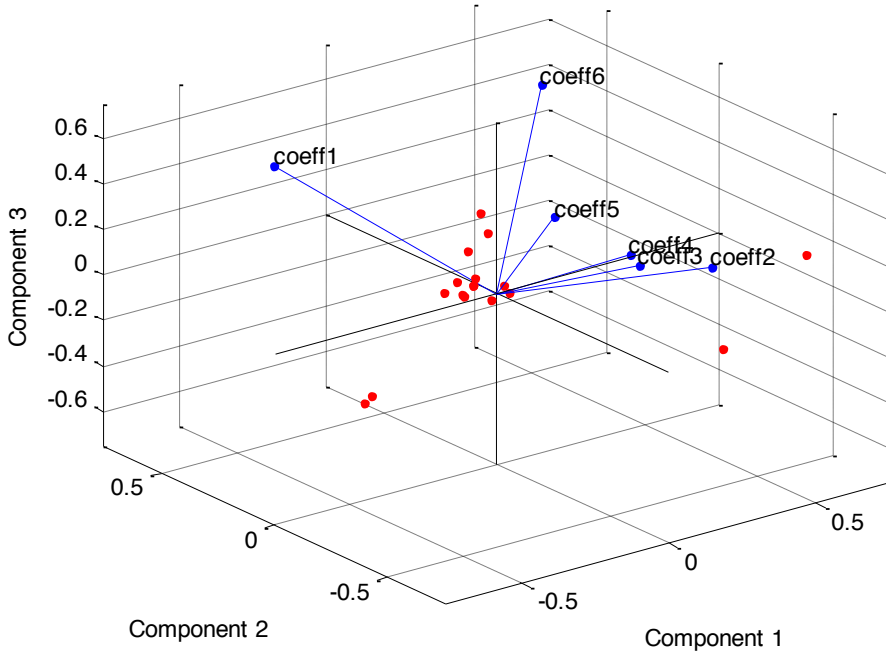
- The Support Vector Machine (SVM) is a technique for classification and regression. Originally the SVM was devised for binary classification, or classifying data into two types. Generalization when there are more than two classes is relatively straightforward.
- For linearly separable data, SVM finds optimal decision boundary using a linear decision surface. When working with non-linearly separable data in the original space, SVM maps the patterns to a higher dimensional feature space in which the transformed data becomes linearly separable. This conversion can be done using kernel function, and the commonly used kernels functions are listed below:

Name of Kernel Function	Definition
Linear	$K(\mathbf{u},\mathbf{v}) = \mathbf{u}^T\mathbf{v}$
Polynomial of degree d	$K(\mathbf{u},\mathbf{v}) = (\mathbf{u}^T\mathbf{v}+1)^d$
Gaussian Radial Basis Function (RBF)	$K(\mathbf{u},\mathbf{v}) = e^{-\frac{1}{2}[(\mathbf{u}-\mathbf{v})^T\Sigma^{-1}(\mathbf{u}-\mathbf{v})]}$
Sigmoid	$K(\mathbf{u},\mathbf{v}) = \tanh[\mathbf{u}^T\mathbf{v}+b]$

- Solution to SVM can be formulated as a Quadratic Programming Problem. It can be easily implemented by most of the popular statistical languages (MATLAB, R, etc.) or packages (LibSVM).

How ANKAM uses Support Vector Machine to improve Models for Finance

Figure 2. Fund Selection using Support Vector Machine



ANAKM PCA-SVM Fund Classifier

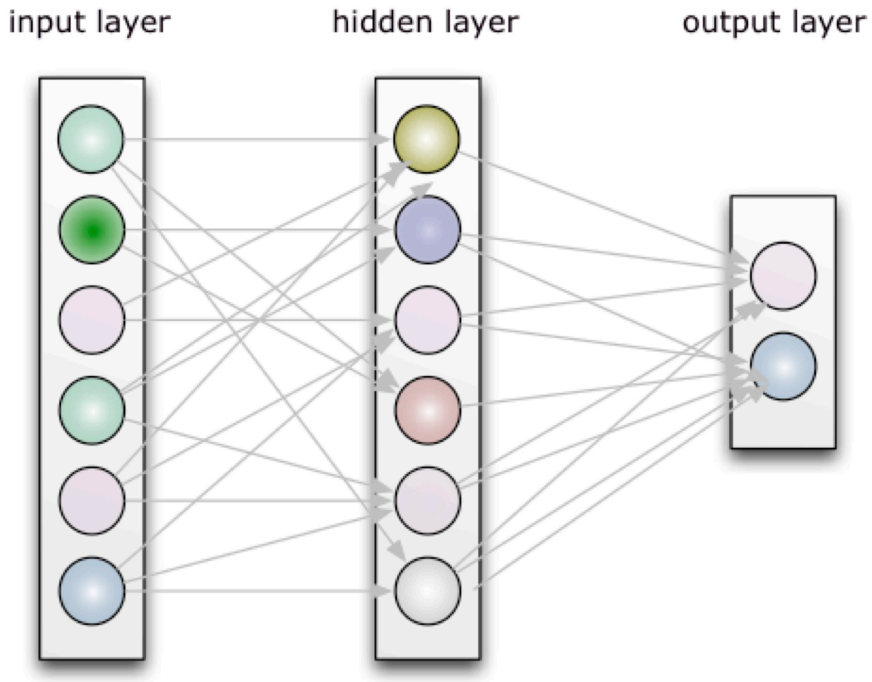
The objective is to classify a group of investment funds into several categories. We first extract a set of feature vectors. We then perform a multi-linear regression. Further, a Principal Component Analysis (PCA) is performed on the coefficient matrix for dimensionality reduction. We then apply SVM to the PCA transformed data to perform fund classification.

What is a Neural Network (NN)?

- Traditional NN uses a feedforward network structure and usually has only one layer. Compared with Deep Neural Network, its structure is simpler and the training is less computationally intensive.

- NN is useful when we have abundance of labeled data but without the knowledge of the underlying mapping function that generates the output. It also shines when data sets are noisy or containing missing variables.

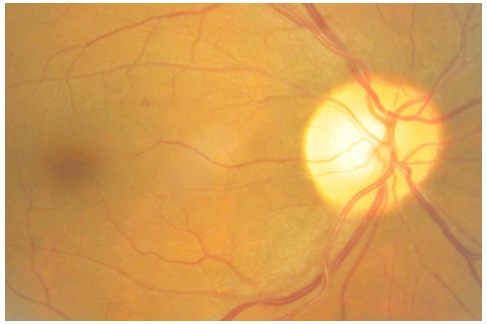
- To train a NN, we first acquire labeled inputs (as high-dimensional vector) and outputs. We then design the structure of the network, such as number of layers and number of neurons in each layer. The formal training process starts with random initialization and feedforward and backpropgation.



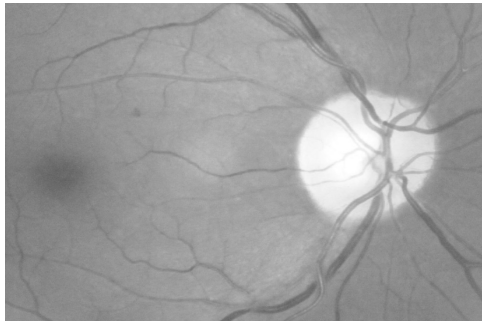
How ANKAM use Neural Net for Diabetic Retinopathy diagnosis

Original Image

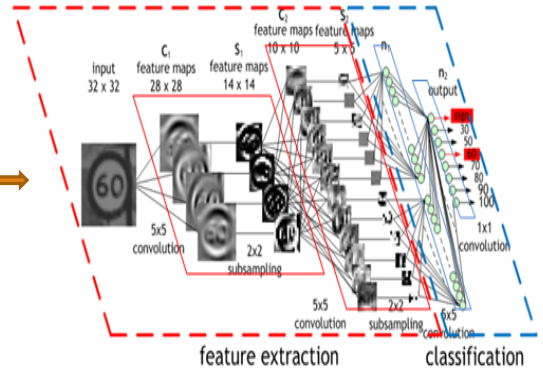
After Adaptive Equalization



Cropped Image



Extracted Green Channel



Run through Neural Net

Diagnosis

Load Retinal Scan

DICOM based storage and exchange of medical image-data

Image Preprocessing

Gaussian Filters, Adaptive Contrast Equalization, etc.

Unsupervised feature extraction

Using CNNs trained on large amounts of data

Supervised learning

Logistic regression, Decision Forest, etc.

Output decision

Give rating based on severity

Introduction to Deep Learning

- What is Deep Learning
  - It is a paradigm in machine learning in which researchers train computer algorithms to spot meaningful patterns by showing them lots of data, rather than explicitly program the rules
  - Neural Nets that mimic the human brain
  - Deep architecture to enable predictions or classifications with unseen accuracy
  - It enables universal classification: the same model can classify languages, images, videos, audios
- Advantages of Deep Learning
  - Scalable: can scale to billions of parameters to learn complex concepts
  - Fast: Model training is fast and a trained model makes online prediction for unseen inputs
  - Unified: it brings a unified approach to data-driven knowledge discovery

Deep Learning for Computer Vision –Natural Scene Parsing

Figure 4. Deep Learning for Scene Understanding (simulated)

Richard Socher [2] proposed a max-margin structure prediction architecture based on recursive neural networks that can successfully recover such structure both in complex scene images as well as sentences. The method outperforms Gist descriptors for scene classification by 4%. We implemented his approach using our research framework and is able to replicate his result.

In the image on the right, the algorithm automatically detects different regions (e.g., sky, tree, etc.), which are colored differently.

The Merck Molecular Activity Challenge

- When developing new medicines it is important to identify molecules that are highly active toward their intended targets but not toward other targets that might cause side effects.

- The objective of this competition was to identify the best techniques for predicting biological activities of different molecules, given numerical descriptors generated from their chemical structures.

- The challenge was based on 15 molecular activity data sets, each for a biologically relevant target. Each row in the dataset corresponds to a molecule and contains descriptors derived from that molecule's chemical structure.

Merck Challenge, Deep Learning and Virtual Screening

- The first prize in Merck Kaggle challenge went to a team of academics who used Deep Neural Networks trained by GPUs.

- After the Merck Kaggle challenge was won by Deep Learning, there has been a heightened interest in applying Deep learning to drug design.

- Unterthiner et al [3] evaluated several techniques including SVM and Deep Learning, on a large database – the CHeMBl database, which has 13 million compound descriptors, 1.3 million compounds, and 5,000 drug targets.

- In terms of accuracy, as measured by AUC (area under the curve), Deep Learning and SVM rated very high compared to conventional techniques like logistic regression or commercial programs like “Pipeline Pilot”.

Virtual Screening, Binding Affinity and Scoring Functions

- Virtual Screening (VS) is a computational technique used in drug discovery to search libraries of small molecules in order to identify those structures which are most likely to bind to a drug target, typically a protein receptor or enzyme. The end result is the reduction in the number of subsequent in- vitro and in-vivo experiments required.
- The goal of VS is to identify a small number of compounds that are far more likely to bind to a given antibiotic drug target than compounds selected at random.
- Scoring functions are used to predict the strength of association or binding affinity between two molecules, such as, a small organic compound (drug) & the drug's biological target (a protein receptor)

Virtual Screening, Neural Networks & Random Forests

- Unfortunately, these scoring functions have had drawbacks since they have many false positives and negatives. **It is here that Machine Learning is proving to be most powerful.**
- Neural Network based scoring functions are emerging as a powerful alternative. E.g., **NNScore**[4] and **NNScore2** [5]
- Random Forests are also used – e.g., **RFScore**[6] based on Binding Affinity data for Protein-Ligand complexes to implicitly capture binding effects that are hard to model explicitly(Ballester and Mitchell)

Structure Activity Relationship (SAR) and SVM

- Structure–Activity Relationship (SAR) analysis is used to reduce the search for new drugs. The aim of SAR analysis is to discover rules that successfully predict the activity of a previously unseen compound based on its physico-chemical descriptors.
- Standard statistical approaches are less useful for SAR problems where
  - data sets contain few compounds and many descriptors
  - the relationship between structure and activity is highly non-linear
- Machines Learning Techniques have the power to tackle these issues. Burbidge [7] shows the power of Support Vector Machines in SAR analysis
  - **The Problem**: To predict the inhibition of an enzyme, dihydrofolate reductase by pyrimidines.
  - **Goal**: The task is to learn the relationship $d_n > d_m$, which states that drug *n* has a higher activity than drug *m*.

About ANKAM

- Ankam Private Limited is founded and managed by Saurabh Singal.
- Saurabh was the manager of the Indea Ankam Fund, a systematic equities hedge fund that used quantitative strategies with great success. His last assignment prior to running the Ankam Fund was with Deutsche Bank.
- The team he led built world class data-intensive analytics. He has over 20 years' experience in portfolio management, securities trading and quantitative research. He worked on the Derivatives Trading and Structuring desks of Credit Suisse First Boston, Merrill Lynch and Deutsche Bank in Tokyo. He holds a B. Tech. in Computer Science from IIT Delhi and an M.S. in Computational Finance from Carnegie Mellon University.
- Saurabh has applied Statistical Learning and Computational Models to problems in Finance and is now looking to apply machine learning in the field of Healthcare with current focus on personalized Medicine.

The ANKAM Competitive Advantage

- Strong infrastructure and knowledge base on Machine Learning
- Strong Research Capability
- Expertise in Finance, Mathematics and Computer Science
- Affinity with leading Indian hospitals
- Strong relationship with leading doctors and surgeons
- Deep connections with leading researchers and academic institutions

References

[1]. Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[2]. Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 129-136).

[3]. Unterthiner, T., Mayr, A., ünter Klambauer, G., Steijaert, M., Wegner, J. K., Ceulemans, H., & Hochreiter, S. (2014). Deep learning as an opportunity in virtual screening. In *Deep Learning and Representation Learning Workshop, NIPS*.

[4]. Durrant, J. D., & McCammon, J. A. (2010). NNScore: A neural-network-based scoring function for the characterization of protein− ligand complexes. *Journal of chemical information and modeling*, *50*(10), 1865-1871.

[5]. Durrant, J. D., & McCammon, J. A. (2011). NNScore 2.0: a neural-network receptor–ligand scoring function. *Journal of chemical information and modeling*,*51*(11), 2897-2903.

[6]. Ballester, P. J., & Mitchell, J. B. (2010). A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics*, *26*(9), 1169-1175.

[7]. Burbidge, R., Trotter, M., Buxton, B., & Holden, S. (2001). Drug design by machine learning: support vector machines for pharmaceutical data analysis.*Computers & chemistry*, *26*(1), 5-14.