JYOTI NIVAS COLLEGE
POST GRADUATE CENTRE

DEPARTMENT OF MCA

TECH-ON-TAP (E-JOURNAL)

MACHINE LEARNING APPLICATIONS

ISSUE 6

FEBRUARY 2021

INDEX

INSURANCE FRAUD DETECTION

SHRESTA.R, REG NO :182MCA34

Problem Statement:

Detection of an insurance fraud is a challenging problem for the insurance industry. Hard insurance fraud is defined as when people intentionally fake an accident. When a person has a valid insurance claim but falsifies part of the claim is known as soft insurance fraud.

Financial fraud has been a big concern for many organizations across industries; billions of dollars are lost yearly because of this fraud. The insurance industries consist of more than thousand companies in worldwide. And collect more than one trillions of dollars premiums in each year. When a person or entity make false insurance claims in order to obtain compensation or benefits to which they are not entitled is known as an insurance fraud. The total cost of an insurance fraud is estimated to be more than forty billions of dollars. So detection of an insurance fraud is a challenging problem for the insurance industry. The traditional approach for fraud detection is based on developing heuristics around fraud indicator. The auto\vehicle insurance fraud is the most prominent type of insurance fraud, which can be done by fake accident claim focusing on detecting the auto\vehicle fraud by using, machine learning technique. Also, the performance will be compared by calculation of confusion matrix. This can help to calculate accuracy, precision, and recall.

Parameters Used

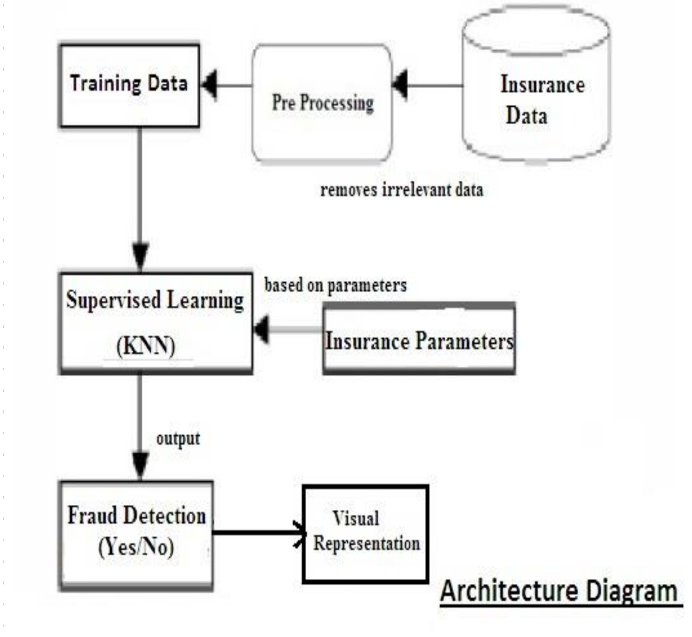
- ✧ Claim number
- ✧ Policy number
- ✧ Claim occurrence start date
- ✧ Claim occurrence start time
- ✧ Claim open date
- ✧ Claim loss date
- ✧ Claim event location name
- ✧ Claim amount
- ✧ Policy premium
- ✧ Part market cost
- ✧ Claim on vehicle
- ✧ Count of customer communication
- ✧ Are claim document submitted
- ✧ Claim event witness name

Then transform this raw data into the transformed data set, and fed into the classification algorithms like decision tree, random forest, and naive bayes. The transformed data set contain following attributes.

- Difference between claim occurrence date and claim report date
- Difference between claim report date and claim open date
- Difference between policy effective and claim occurrence date
- Are claim document submitted
- Part cost difference
- Credit rating
- Part cost difference
- Credit rating
- Policy premium
- Claim occurrence start time
- Count of customer communication
- Claim event location name

By using these attribute, check the each details about the claims. In case of normal claims, the gap claim occurrence date and claim report date is less than seven. The all document are submitted with proof. The gap between policy effective date and claim occurrence date is less than five days. Also check the claims on same vehicle during all policy periods.

DESIGN:



Architecture Diagram

METHODOLOGY SUPERVISED LEARNING

Classification Rules

Classification is a process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.

Example: Suppose the sales manager of All Electronics want to classify a large set of items in the store, based on three kinds of responses to sales campaign: good response, mild response and no response. The model for each of these three classes is derived based on the descriptive features of the items, such as price, brand, place made, type and category. The resulting classification should maximally distinguish each class from the others, presenting an organized picture of the data set.

Introduction to k Nearest Neighbours algorithm

In machine learning, k Nearest Neighbours or kNN is the simplest of all machine learning algorithms. It is a non-parametric algorithm used for classification and regression tasks. Non-parametric means there is no assumption required for data distribution. So, kNN does not require any underlying assumption to be made. In both classification and regression tasks, the input consists of the k closest training examples in the feature space. The output depends upon whether kNN is used for classification or regression purposes.

- In kNN classification, the output is a class membership. The given data point is classified based on the majority of type of its neighbours. The data point is assigned to the most frequent class among its k nearest neighbours. Usually k is a small positive integer. If k=1, then the data point is simply assigned to the class of that single nearest neighbour.

- In kNN regression, the output is simply some property value for the object. This value is the average of the values of k nearest neighbours.

kNN is a type of instance-based learning or lazy learning. Lazy learning means it does not require any training data points for model generation. All training data will be used in the testing phase. This makes training faster and testing slower and costlier. So, the testing phase requires more time and memory resources.

In kNN, the neighbours are taken from a set of objects for which the class or the object property value is known. This can be thought of as the training set for the kNN algorithm, though no explicit training step is required. In both classification and regression kNN algorithm, we can assign weight to the contributions of the neighbours. So, nearest neighbours contribute more to the average than the more distant ones.

HUMAN MOOD PREDICTION AND TO CONVINCE A PERSON TO EAT HEALTHY FOOD

ZAINAB FATHIMA K Z (182MCA28)

Problem Statement

A system to convince a human being for eating the given food: Human being tend to like some variety of food while dislike other which is considered to be a part of human mood. The problem is to develop a machine learning technology which can understand the type of mood and based on that convince oneself to do what one does not like to do. Here we force a human being to eat a kind of food which one does not like.

Parameters for the given problem

Input: A constant analyser: a sensor (mobile phones), wearable sensors(fitbit& Jawbone), social media, ubiquitous sensing, ambient intelligence(sensor on every device where user lives), google history, application using activities which determine mood(chiku-journal diary/mood tracker)

Data to knowledge: behavioural markers, sensemaking hierarchical network.

Output: clinical states(anxiety, happy) and draw the concluded results for convincing (like give convincing statements in good mood).

Methodology suitable to do so

Transformation tools, regression classification, domain knowledge, brainstorming, slow feature analysis and stacked autoencoders, ecological momentary assessment (EMA).

Algorithmic Steps:

1. Raw Sensor Data
   It mainly focuses on cleansing, extracting only behavioural data.
2. Feature Extraction: Data to information
   Transforming the extracted data into features. Features are constructs measured by, and proximal to, the sensor data. One approach is to use domain knowledge or brainstorming to inject human intelligence for feature construction. Features can also be extracted using statistical models such as slow feature analysis and stacked autoencoders.
3. Behavioural Markers: information to knowledge
   Behavioural markers are higher-level features, reflecting behaviors, cognitions, and emotions that are measured using low-level features and sensor data. They are most

commonly developed machine learning and data-mining methods to uncover which features and sensor data are useful in detecting the marker.

4. Clinical targets: limited sets of features have been modestly successful at predicting clinical targets. Although they might be symptoms undetectable and also this personal sensing may uncover other predictors that have not been considered till date.
5. Processing:(I feel) A network populated with knowledge(hierarchical structure) may be used to draw out statements apt for convincing the output mood.

Machine learning used: An autonomous prediction machine useful to think of a mental-health sensing platform as a social machine in which the quality of prediction is ensured through a shared endeavour. Here regression classification technique will differentiate the sensed data.

References:

- Asselbergs J, Ruwaard J, Ejdys M, Schrader N, Sijbrandij M, Riper H. 2016. Mobile phone-based unobtrusive ecological momentary assessment of day-to-day mood: an explorative study. J. Med. Internet Res 18:e72.
- David C. Mohr, Mi Zhang, and Stephen M. Schueller, Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning

PREDICTING COOKING MENU

SHIVAVARSHNI K (18MCA18)

Abstract:

Cooking is a basic need for people to prepare food for consumption. Choosing what to cook for a person is another situation to overthink. If the person has a heath issue he/she needs more care and to be provided with healthy prepared dishes. Machine learning can be applied to this situation that is the ingredients can be sent as the parameter and it will give a solution what to be prepared. This method can also be helpful to the people who cook every day and always has a question "What to Cook?"

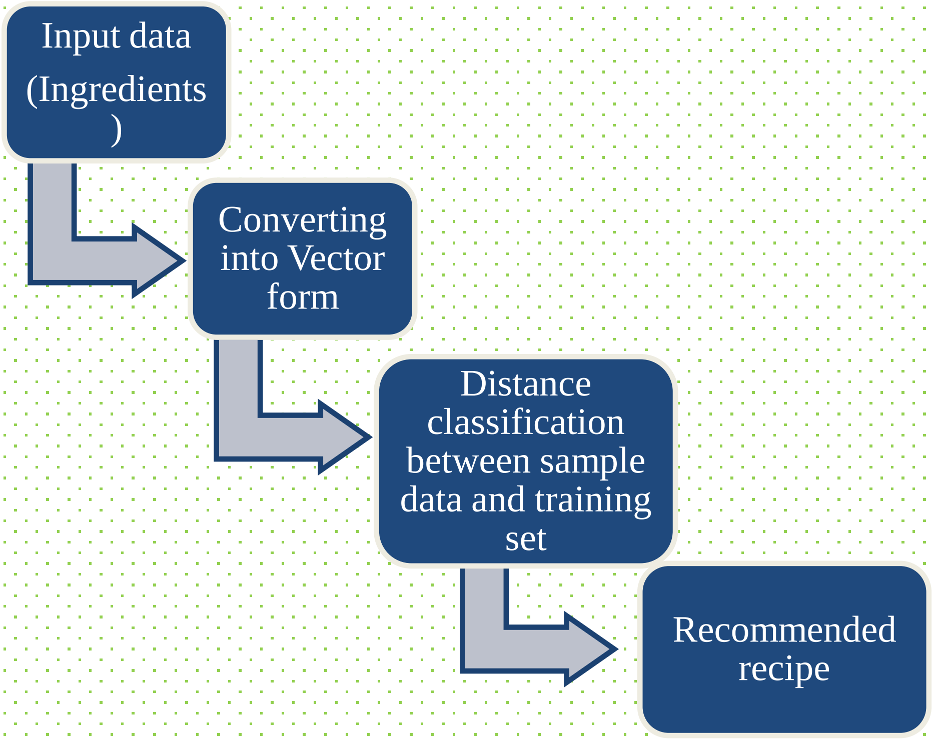
Parameters:

- ✓ Ingredients
- ✓ Person's heath data
- ✓ Type(Veg/NonVeg)
- ✓ Number of dishes
- ✓ Number of Consumers
- ✓

Methodology:

K-nearest neighbors classification: input text is converted into a vector format and classified with the trained data and appropriate class is defined where the matched recipes are recommended to the user. KNN will be better since the class classification and distance is calculated easily and much number of dishes can be predicted.

Algorithmic Steps:


Input data (Ingredients)
Converting into Vector form
Distance classification between sample data and training set
Recommended recipe

Conclusion:

Using ML many day to day problems can be solved. As cooking is a daily requirement I would like to try to solve this issue using ML.

HAIRSTYLE RECOMMENDATION

DINAMANI R (182MCA33)

Introduction:

Identifying human face shape is the first and the most vital process prior to choosing the right hairstyle to wear on according to guidelines from hairstyle experts, especially for women. This work presents a novel framework for a hairstyle recommender system that is based on face shape classifier. This framework enables an automatic hairstyle recommendation with a single face image. This has a direct impact on beauty industry service providers. It can simulate how the user looks like when she is wearing the chosen hairstyle recommended by the expert system. The model used in this framework is based on Support Vector Machine. The framework is evaluated on hand-crafted, deep-learned (VGG-face) features and VGG-face fine-tuned version for the face shape classification task.

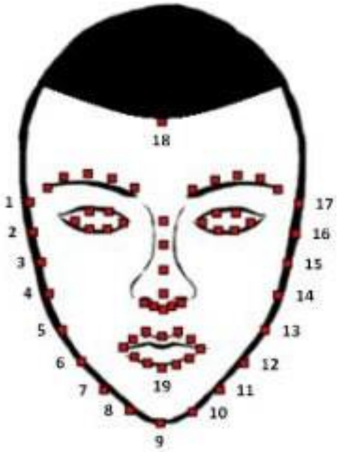
Facial Feature Extraction for Face Shape Classification:

Face shapes: Round shape, Oval shape, Oblong shape, Heart shape, Square shape

Hand-Crafted Feature:

The face region is required to be identified first in order to determine the face shape. To identify the face region, land-mark localization technique called Active Appearance Model(AAM) in conjunction with color-based skin segmentation(Mustafa, 2007) is utilized. AMM is a technique to obtain vital points on face images. In addition to AAM, color-based skin segmentation can define the point indicating the hairline that separates the hair and forehead. Then, these points are used to calculate the geometric features to represent face shape as hand- crafted features.

- Ratio of the height of a face to the width
- Ratio of the distance between both sides of the jaws to the width of the face
- Ratio of the distance between the chin and the bottom of the mouth to the distance between both sides of the jaws
- Angle between the straight line from the boundary of the face at a considered point to the chin point and the horizontal vector



Facial Landmark Points

The operation of hairstyle recommender system consists of three main steps as follows:

- Getting an image of the user's from a camera
- Classifying the face into the right shape, and
- Hairstyle retrieval based on the face shape and superimposing it on her face

Four main attributes which are very important to de-scribing hairstyle as follows:

- Length={pixie, short, mid-length, long};
- Style={straight, wavy, mix}
- Bang={none, blunt, side-swept}
- Layered={non-slide, slide}

SVM :

SVM is one of the most popular machine learning techniques. It is a non-parametric model. Compared to other popular learning techniques in computer vision such as Artificial Neural Network, it has a distinct ability to prevent over fitting problem. Moreover, it can be defined by a convex optimization problem. However, it requires high human intervention to determine the optimum parameter settings such as regularization parameter, choice of kernel and its parameter.

A hairstyle recommendation system based on face shape classification approach. Once a face shape is identified, the system will recommend a set of hairstyles based on guidelines from an expert. The classification model used in this system is based on SVM-RBF. Propose an approach to improve the performance of the classifier from that of using only individual features.

- The geometric feature extraction procedure, the current approach utilizes the color-based face segmentation to identify the top point of forehead. This did not work when the user wore fringe as it covered her forehead. It would1045be better if there is a model that can predict the top point of the forehead with more stability.

In terms of feature improvement, investigating several pre-trained models as well as trying different layers in the1055pre-trained model are keys to explore more appropriate deep-learned features for face shape classification. There is a re-search that proposes a systematic framework to select an appropriate layer of pre-trained deep CNN-based face recognition for face attribute prediction. The intermediate layers of the model might keep more physical facial characteristics than the last two layers to.

ENHANCING THE LEARNING PROCESS

JYOTSANA R JAIN (182MCA52)

PROBLEM STATEMENT:

In the pandemic situation, to ensure the student’s education is not being hampered, everything around is being virtualized to continues the learning process.

Combating the current scenario, it is challenging to multitask wherein we need to concentrate to understand the concept and jot down the important points at the same time ensuring they are not missed as it would subsequently help to recollect and revise the same.

A solution to make the learning process better, the ML (machine) would convert the audio into text notes so we need not multitask and can completely focus on understanding the concept well and revise the concept with the help of the auto text being captured.

ATTRIBUTES / PARAMETERS:

- ✓ Imprecise Interpretation – VUI’s (voice user interface)
- ✓ Time – diversity of voice patterns that human possesses and precise in pronunciation
- ✓ Accents – comprehend dialects
- ✓ Background noise and loudness – loud environment
- ✓ Amplitude – refers to the maximum displacement of the air molecules from the rest position
- ✓ Crest and Trough – the crest is the highest point in the wave whereas trough is the lowest point
- ✓ Wavelength – the distance between the two successive crests and troughs
- ✓ Cycle – every audio signal traverses in the form of cycles. One complete upward movement and downward movement of the signal forms a cycle
- ✓ Frequency – refers to how fast a signal is changing over a period of time

METHODOLOGY:

There are three learning methods in machine learning- supervised learning, unsupervised learning and reinforcement learning methods.
Supervised learning is a learning method that maps an input to an output based on input-output pairs. And in unsupervised learning we do not know the targeted outcome and we should train the model in order to get desired outcome.

PROBABILISTIC GRAPHICAL MODELS: Probabilistic graphical models use a graph- based representation as the foundation for encoding a distribution over a multi-dimensional space and a graph that is a compact or factorized representation of a set of independences that hold in the specific distribution.

HIDDEN MARKOV MODEL (HMM): Modern general-purpose speech recognition systems are based on Hidden Markov Models. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal.

In a short time-scale (e.g., 10 milliseconds), speech can be approximated as a stationary process. they can be trained automatically and are simple and computationally feasible to use.

Data set: Supervised learning. Each clip contains one of the 30 different voice commands spoken by thousands of different subjects. It also contains amplitude, wavelength, cycle, frequency etc.

ALGORITHM:

HIDDEN MARKOV MODEL (HMM) - The parameter learning task in HMMs is to find, given an output sequence or a set of such sequences, the best set of state transition and emission probabilities. The task is usually to derive the maximum likelihood estimate of the parameters of the HMM given the set of output sequences. The hidden Markov model will tend to have in each state a statistical distribution that is a mixture of diagonal covariance Gaussians, which will give a likelihood for each observed vector. Each word, or (for more general speech recognition systems), each phoneme, will have a different output distribution; a hidden Markov model for a sequence of words or phonemes is made by concatenating the individual trained hidden Markov models for the separate words and phonemes.

Step – 1: import libraries – libROSA and SciPy are used for processing audio signals

Step – 2: Data exploration and Visualization – import of dataset, helps to analyse and understand the data as well as pre-processing steps in a better way. Visualization of audio signals are in time series domain

Step – 3: sampling and resampling rate – sampling and resampling rate of the audio signals based on the speech related frequencies. (ex. Sampling rate to be 16000 Hz to resampling rate to be 8000 Hz)

Step – 4: duration of recordings – the distribution of the duration of the recordings

Step – 5: pre-processing of audio waves on 2 aspects, resampling and removing of shorter commands less than 1 second.

Step – 6: split into train and validation set – train the model on 80% of the data and validate on the remaining 20%

Step – 7: input – main module – max pooling – dense – output (model architecture).

Step – 8: Diagnostic plot – data visualization to understand the performance of the model over a period of time.

Step – 9: load the model and predict the suitable text for the audio file on the validation data.