IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 83-87

A Survey of Feature Extraction and Feature engineering In Data Mining

Nithya C[1] , Saravanan V[2]

Department Of Computer Science, Hindusthan College Of Arts And Science, Coimbatore, India.
Department Of Information Technology, Hindusthan College Of Arts And Science, Coimbatore, India.

Abstract:*Data mining is the process of discovering interesting knowledge patterns from large amount of data stored in database. It is an essential process where the intelligent techniques (i.e., machine learning, artificial intelligence, etc) are used to extract the data patterns (i.e., features). The aim of data mining process is to extract the useful information from dataset and transform it into understandable structure for future use. Feature extraction is the process of extracting the relevant features from large database for dimensionality reduction. Feature extraction is a key process to reduce the dimensionality of medical dataset for efficient disease prediction. The feature extraction technique removes irrelevant features to acquire higher prediction accuracy during disease diagnosis. Few research works are developed to extract the relevant features from dataset using different data mining techniques. Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive. The need for manual feature engineering can be obviated by automated feature learning.*

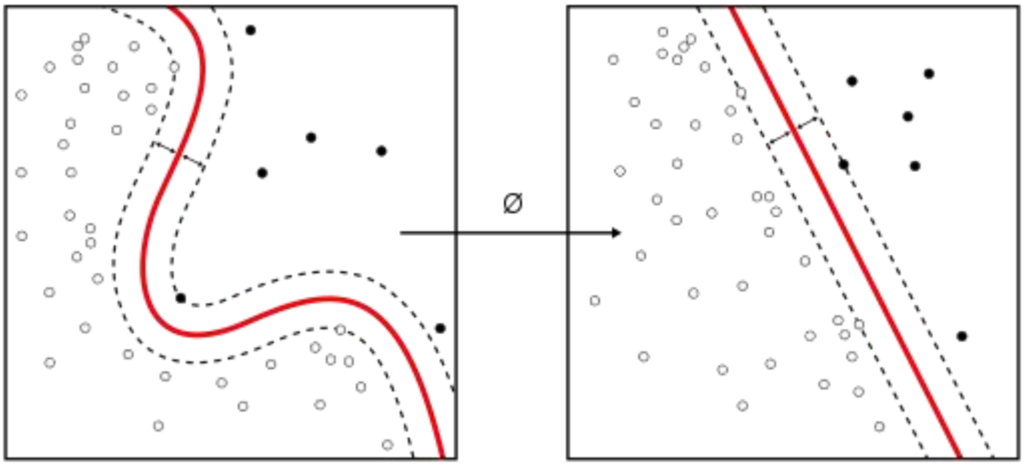
Keywords: *Data Mining, Feature Extraction, Machine Learning, Feature Engineering.*

I. Introduction:

Data mining techniques has the ability to provide the set of useful rules for performing the tasks. Feature extraction is an essential process for addressing the machine learning problems. Features extraction is an essential one for the implementation of decision support system as it identifies abnormal one through selecting the essential features. The feature extraction techniques aimed on global structure for dimensionality reduction. Feature extraction is used to encode the high dimensional data into low dimensional space. The feature extraction results are enhanced by constructing set of application-dependent features called feature engineering. Feature engineering is an informal topic, but it is considered essential in applied machine learning. Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.Feature engineering is the process of employing the domain knowledge of data to create the features for efficient machine learning algorithms performance. Feature engineering is primary one for machine learning application.

Machine Learning Techniques in Data Mining:

Machine learning is a subset of artificial intelligence in the field of computer science that often uses statistical techniques to give computers the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.



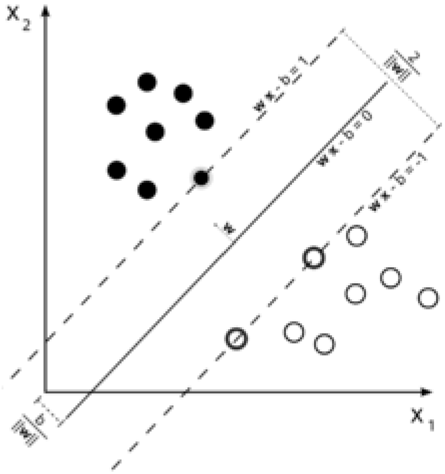
Kernel machines are used to compute non-linearly separable functions into a higher dimension linearly separable function.

Machine learning tasks

Machine learning tasks are typically classified into two broad categories, depending on whether there is a learning "signal" or "feedback" available to a learning system:

- Supervised learning: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. As special cases, the input signal can be only partially available, or restricted to special feedback:
  - Semi-supervised learning: the computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing.
  - Active learning: the computer can only obtain training labels for a limited set of instances (based on a budget), and also has to optimize its choice of objects to acquire labels for. When used interactively, these can be presented to the user for labeling.
  - Reinforcement learning: training data (in form of rewards and punishments) is given only as feedback to the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.[1]
- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

Machine learning applications



A support vector machine is a classifier that divides its input space into two regions, separated by a linear boundary. Here, it has learned to distinguish black and white circles.

Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system:[1]

- In classification, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised manner. Spam filtering is an example of classification, where the inputs are email (or other) messages and the classes are "spam" and "not spam".
- In regression, also a supervised problem, the outputs are continuous rather than discrete.
- In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
- Density estimation finds the distribution of inputs in some space.
- Dimensionality reduction simplifies inputs by mapping them into a lower-dimensional space. Topic modeling is a related problem, where a program is given a list of human language documents and is tasked with finding out which documents cover similar topics.

Feature Extraction

Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes.

Feature extraction technique is used to extracts a subset of new features from the original feature set by means of some functional mapping by keeping as much information in the data as possible. The following methods are commonly used for the feature extraction [3]

A. Principal Component Analysis

Principal Component Analysis (PCA) is the most popular statistical method. This method extracts a lower dimensional space by analyzing the covariance structure of multivariate statistical observations [2].The computation of the PCA transformation matrix S is given as

$$S = \left(\sum_{i=1}^{n} (Y_i - m)(Y_i - m)^T \right) \qquad (1)$$

where
n is the number of instances
Yi is the i-th instance
M is the mean vector of the input data

B. Linear Discriminant Analysis

Linear Discriminant Analysis
(LDA) technique mainly projects the high-dimensional data into lower dimensional
space. LDA aims to maximize the between-class distance and minimize the within-class distance in the dimensionality-reduced space [2]. The LDA is computed as

$$f(X) = trace\left((X^T S_w X)^{-1} (X^T S_b X) \right) \qquad (2)$$

where
Sb is the between-class matrix
SW is the within-class matrix

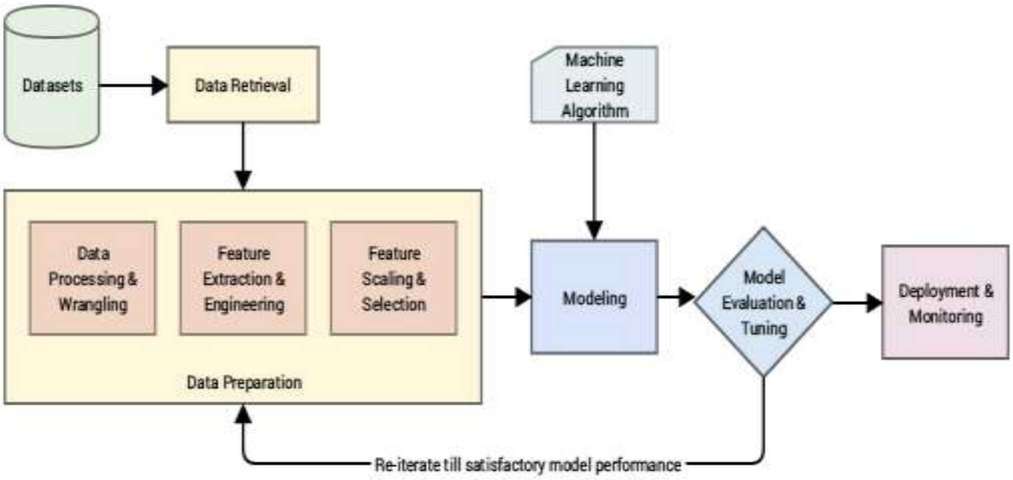
$$S_b = \frac{1}{n} \sum_{i=1}^{m} k_i \, (c_i - c)(c_i - c)^T$$

$$S_w = \frac{1}{n} \sum_{i=1}^{m} \sum_{x \in X_i} (x - c_i)(x - c_i)^T$$

where,
Xi is the index set of ith class
Ci is the mean vector of ith class.

Features Engineering:

A feature is an attribute or property shared by all of the independent units on which analysis or prediction is to be done. Any attribute could be a feature, as long as it is useful to the model. The purpose of a feature, other than being an attribute, would be much easier to understand in the context of a problem. A feature is a characteristic that might help when solving the problem[4].


Datasets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

A standard machine learning pipeline

Importance of features

The features in your data are important to the predictive models you use and will influence the results you are going to achieve. The quality and quantity of the features will have great influence on whether the model is good or not[5].

You could say the better the features are, the better the result is. This isn't entirely true, because the results achieved also depend on the model and the data, not just the chosen features. That said, choosing the right features is still very important. Better features can produce simpler and more flexible models, and they often yield better results[4].

The algorithms we used are very standard for Kagglers. […] We spent most of our efforts in feature engineering. [...] We were also very careful to discard features likely to expose us to the risk of over-fitting our model.

— Xavier Conort, "Q&A with Xavier Conort"[6]

…some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used.

— Pedro Domingos, "A Few Useful Things to Know about Machine Learning"[7]

The process of feature engineering[8]:

1. Brainstorming or Testing features;
2. Deciding what features to create;
3. Creating features;
4. Checking how the features work with your model;
5. Improving your features if needed;
6. Go back to brainstorming/creating more features until the work is done.

Automated Feature Engineering

Automation of feature engineering has become an emerging topic of research in academia. In 2015, researchers at MIT presented the Deep Feature Synthesis algorithm and demonstrated its effectiveness in online data science competitions where it beat 615 of 906 human teams[9][10]. Deep Feature Synthesis is available as an open source library called Featuretools. That work was followed by other researchers including IBM's OneBM [11] and Berkeley's ExploreKit[12]. The researchers at IBM state that feature engineering automation "helps data scientists reduce data exploration time allowing them to try and error many ideas in short time. On the other hand, it enables non-experts, who are not familiar with data science, to quickly extract value from their data with a little effort, time and cost."

II. Conclusion:

In this paper, a survey is carried out about Feature Extraction and Feature Engineering in data mining to extract the new set of features efficiently.Mainy feature extraction algorithms proposed by different researchers are discussed and the issues present in the existing algorithm were identified. Hence, the future work is to overcome the issues and to propose a new feature extraction and future engineering algorithm which will extract the new set of features.

References:

[1]. Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Springer, ISBN 0-387-31073-8

[2]. Daniel Engel, Lars Hüttenberger, Bernd Hamann, “A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization”, LNCS Springer, 2014, pp. 1-16.

[3]. Khalid, Samina, Khalil Tehmina, NasreenShamila, “A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning”, IEEE Science and Information Conference, 2014, pp. 372-378.

[4]. "Feature Engineering: How to transform variables and create new ones?". Analytics Vidhya. 2015-03-12. Retrieved 2015-11-12.

[5]. "Discover Feature Engineering, How to Engineer Features and How to Get Good at It - Machine Learning Mastery". Machine Learning Mastery. Retrieved 2015-11-11.

[6]. "Discover Feature Engineering, How to Engineer Features and How to Get Good at It - Machine Learning Mastery". Machine Learning Mastery.Retrieved 2015-11-11.

[7]. http://blog.kaggle.com/2013/04/10/qa-with-xavier-conort/

[8]. Domingos, Pedro. "A Few Useful Things to Know about Machine Learning" (PDF). Retrieved 12 November 2015.

[9]. "Feature engineering and selection" (PDF). Alexandre Bouchard-Côté. Retrieved 12 November 2015.

[10]. https://news.mit.edu/2015/automating-big-data-analysis-1016

[11]. https://dai.lids.mit.edu/wp-content/uploads/2017/10/DSAA_DSM_2015.pdf

[12]. https://arxiv.org/pdf/1706.00327.pdf

[13]. https://people.eecs.berkeley.edu/~dawnsong/papers/icdm-2016.pdf

[14]. https://techcrunch.com/2017/11/30/h20-ais-snares-40m-series-c-investment-led-by-wells-fargo-and-nvidia/

[15]. https://techcrunch.com/2018/02/22/feature-labs-launches-out-of-mit-to-speed-up-building-machine-learning-algorithms/