IFERP
connecting engineers... developing research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 8, Issue 9, September 2021

Significance of Feature Selection and Reduction in Hybrid Classification of High Dimensional Data Sets

[1] Preetha R, [2] Dr S Vinila Jinny

[1] Research Scholar, Noorul Islam Centre For Higher Education, Kumaracoil, Thackalai, India
[2] Associate Professor, Computer Science and Engineering Noorul Islam Centre For Higher Education
Kumaracoil, Thackalai, India

***Abstract*----Customer experience has been one of the focus areas for organizations to gain a competitive advantage in the market. The existing literature is full of various knowledge base, models, methodologies, and some paradigm shift ideas. Since Pine and Gilmore was one of the early researchers to address the concept of customer experience followed by Carbone and Hackle in 1999. We have observed that most of the existing literature is more about hospitality, retail, tourism, and service industry though the products and goods studies are also found to be addressed. One major gap area is the Primary or base industry, where there is limited knowledge available including Automobile, metals and mining and others. We have also found varying definition of "Consumer experience" and most of the studies addressing some or part of overall holistic consumer experience. The main reason for the lack of holistic approach seems to arise from multidimensional nature of customer experience phenomenon. In our literature review we have observed regular discipline of cognitive and hedonic as main but there is literature which comprises psychology, human behavior, economics, anthropology, Neurology; sociology, organizational behavior also contributing to explain the consumer experience phenomenon. This multi-disciplinary nature of consumer experience still leaves many areas to be further researched and explored. From this multidisciplinary point of view, we can safely say the holistic consumer experience research is still in its infancy. The literary review also included models, methodologies, and philosophies to build a great customer experience, some sort of guidance for firms and marketing and service managers about strategy, steps, and measurement in building a great customer experience for their respective customers. Some new concept such as co-creation of values, joint working between the firm and industrial customer, experience design is also observed. Purpose of our study was to go through the various stages of evolution of concept of customer experience from 1980 onwards and understand the knowledge and paradigm shifts which happened in this field in last three decades. We were surprised to also come across some of the work in 1960 and mention of customer experience in Adam Smith's work. Finally, as explained by Pine and Gilmore 1998 and 1999, Shaw and Iven's 2002, Voss 2003, Prahalad and Ramaswamy 2004, Meyer and Schwager 2007, customer experience has become numeri uno priority for the organizations and is seen as a true competitive edge in today's crowded market Place. From our study of literature, we found that there are many more areas such as CX measurement , universal CX definition and multidisciplinary holistic studies where further research is required.**

***Index Terms*— Data mining, classification, Feature reduction, PCA, LDA, ANFIS, Differential Evolution, Margin based, K- nearest neighbor, SVM, Naive Bayes, Decision Tree, ANN**

I. INTRODUCTION

Data mining is the process of extracting visible data from data set in such a way that it can be used by various technologies in the future. The biggest steps in data mining are these

- Identify source details
- Picking up data points that need analysis.
- Release relevant data from the data.
- Identify key values in the output data set.
- Interpreting and reporting results

It is one of the most important processes and requires a lot of attention and patience in collecting the data you want and its functionality. Nowadays data are ruling this world because it has the capability to change the face of our world.

For that, the term Data mining has its importance. It provides a variety of effective strategies for extracting useful information from a data group. The rapid growth of computer biology and e-commerce of high-quality data is becoming commonplace. Higher size data mining is a very important problem. High dimensional data sets have the problem of curse of dimensionality and poor generalization ability. There are two approaches to address these challenges:

- Reduce the size of the data set
- Independent methods for data size are used

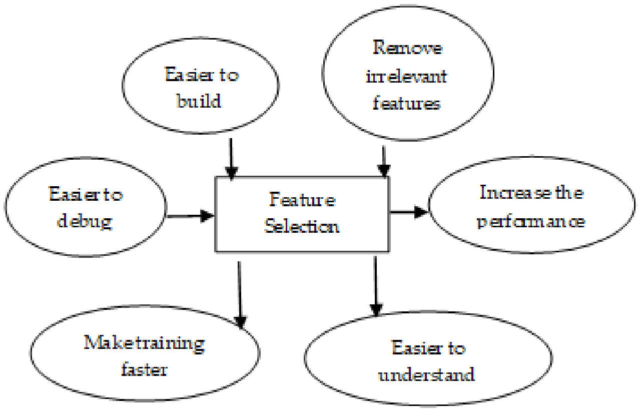
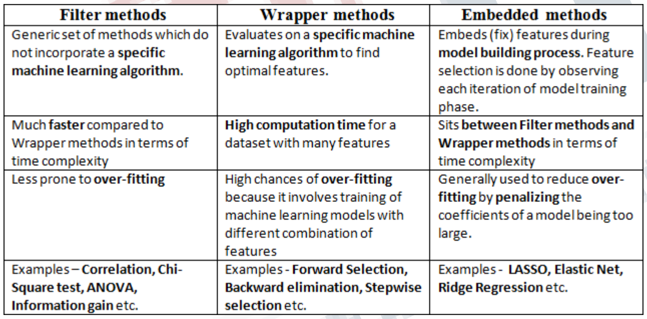


Fig. 1. Feature Selection.

There are two things to reduce size: Select feature and reduce feature. Therefore the selection of features and the reduction of the feature is done as a step to reduce the size before building the partition model. Feature selection is the process of obtaining a subset of real variables, and that subset is used to measure a problem. It is divided into three types:

- Wrapper Methods
- Filter Methods
- Embedded Methods



Filter methods	Wrapper methods	Embedded methods
Generic set of methods which do not incorporate a **specific machine learning algorithm.**	Evaluates on a **specific machine learning algorithm** to find optimal features.	Embeds (fix) features during **model building process.** Feature selection is done by observing each iteration of model training phase.
Much **faster** compared to Wrapper methods in terms of time complexity	**High computation time** for a dataset with many features	Sits **between Filter methods and Wrapper methods** in terms of time complexity
Less prone to **over-fitting**	High chances of **over-fitting** because it involves training of machine learning models with different combination of features	Generally used to reduce **over-fitting** by **penalizing** the coefficients of a model being too large.
Examples – **Correlation, Chi-Square test, ANOVA, Information gain** etc.	Examples - **Forward Selection, Backward elimination, Stepwise selection** etc.	Examples - **LASSO, Elastic Net, Ridge Regression** etc.

Fig. 2. Difference between Filter, Wrapper, and Embedded Methods for Feature Selection

The feature selection process in wrapper methods is based on a certain machine learning algorithm that we are attempting to fit to a given data set. It uses a greedy search strategy, assessing all potential feature combinations against the evaluation criterion. This method generates a large number of models with different subsets of input features, and selects the features that lead to the most efficient model based on performance metrics. Although they can be computationally expensive, these methods are unconcerned with variable types.

Selection criteria selection methods use mathematical techniques to assess the relationship between each input variation and target variable, with scores used to select the input variables (filter) to be included in the model. Filter methods use their combinations and variations depending on determining the value of the features. Algorithm for selecting feature in embedded paths

There are several advantages to each model. While the filtering process is relatively inexpensive as a compact model, the latter is more accurate in class predictions than ever before as it uses differentiation in its selection process. However, the folding model is often overused especially with high size, and the filter is easy to operate

Feature reduction is the process of reducing data from high to low, or one with a smaller size.

Model training we collect a lot of data to help the machine learn better. Having too much unnecessary data can cause the model to be slow. The model can also learn from this irrelevant data and be inaccurate. Hence feature selection is used for the process of reducing the input variable by selecting only relevant data and getting rid of noise in data. We can optimize the model in many ways by using feature selection. Feature selection is classified into supervised and unsupervised. The output class labels does not required in unsupervised feature selection but in supervised feature selection the output class label is required for the feature selection.

A. Need of Data mining

Data mining is a technique for uncovering hidden information in a database. The second type of data analysis is exploratory data analysis, which involves deducing and learning from data. It could be obtaining useful data from a database.

Needs comes from evolution in size of data base. Manual analysis is not possible. Hence automatic analysis required and for that data mining is essential for extracting useful information from the database.

Various data mining technologies re-analyze, group integra- tion, standard deviation etc., which is the basis of data mining. Artificial Intelligence is used to incorporate human thought as processing. Machine learning is a union of mathematics and artificial intelligence.

II. IMPACT OF FEATURE SELECTION ON MEDICAL DATA SETS

A major class of medical problems involves disease diagnosis, which is predicated on different tests performed on the patient. Therefore the use of class models in clinical trials is increasing significantly. Improving the severity of a separate learning segment is better because it is a combination of two or more models to avoid individual shortcomings and achieve high accuracy. A good selection of features can be key to creating a well-balanced and accurate separation. In this paper we studied various feature selection methods and then applied classification algorithms on the selected features and performed a comparison study about the accuracy of the system.

According to statistics from Indian oncology departments, breast cancer is the second leading cause of death in women after all other cancers. The most effective strategy to prevent breast cancer death is to diagnose it early. To identify between benign and malignant tumors early on, an accurate and reliable method is required. A number of studies have been conducted. It demonstrates that breast cancer is occurring at considerably younger ages than in the past. The number of cases of breast cancer in women of all ages is constantly increasing. Breast cancer in young women is often more aggressive than cancer in adults and survival in younger patients especially in advanced stages is lower.

Selection of novel feature and integration learning the algorithm is necessary to increase the accuracy of the breast cancer stage based on health information in order to provide medical practitioners with accurate and reliable prediction results. In studying together we look at multiple divisions and combine multiple classifiers to obtain better predictions or segmental accuracy by combining multiple subtraction releases we find the result that is the best result of any classification. From the various surveys we found that a suitable feature selection method can improve the accuracy of breast cancer. Several methods are applied to detect the presence of cancer within the breast. The selection of margin-based features has a high relevance that reflects the quality of the features. Margins are used to evaluate the fitness of features. In differential evolution distribution factor and the relative feature distribution factor are used to select the most appropriate features in the feature set. In medical field, prediction systems should be accurate and effective. The absence of those systems opened the path for a potentially dangerous future. To get out of this bad predicament, utilize a classification method on a large database. The curse of dimensionality, which typically degrades the efficiency of machine learning algorithms, is one of the most significant obstacles in the study of datasets. In principal component analysis, principal components are generated by using the correlation between the features. In linear discriminant analysis feature reduction performed by minimize variance and maximizing the distance between the means of the classes. In this paper we are that specialize in the impact of various feature reduction technique using same dataset. Feature reduction is an inevitable role in the classification process. It can influence the performance of a classifier to a certain extent.

III. MARGIN BASED FEATURE SELECTION

It captures the relevant aspects of the data from a given collection of features in margin based feature selection (MBFS). It does it by using the margin parameter. A margin is a geometric measure that can be used to assess a classifier's confidence in relation to its choice. In modern machine learning research, margins already play a significant role.

There are two natural ways to define a model line in relation to the law of segregation. Very common type, sample margin, measures medium distance for example and the decision limit caused by to separate. For our study we haves used this simple one which makes the computation more simpler.

After evaluating the margin score among features, margin based feature selection selects the features with the greatest margin score. These characteristics are also used throughout the training and testing phases.

IV. DIFFERENTIAL EVOLUTION BASED FEATURE SELECTION

This is an algorithm for selecting a novel feature based on the use of Differential Evolution (DE) method Proposed DEFS significantly reduces computer costs while simultaneously demonstrating effective performance.

The first step in the algorithm is to generate new human vectors from real people. For each position in the human matrix, the alternating vector is created by first selecting two random vectors then make a weighted difference, adding effect to 3 random (base). The vector changes and falls with the real vector that stays in that position the real matrix. The result of this operation is called the trail vector. The corresponding position is new the number of people will consist of a trail vector (or its modified version) or a specified vector depending on which of them has reached the highest intensity (precision classification).

Round of the lead vector may contain redundant values affects the performance. To overcome such feelings a roulette wheel alignment system is used. In this system a cost estimate is used where the probability of an individual's existence is calculated from the distribution factors associated with each element.

V. PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a consistent and straightforward process of transformation. Karl Pearson introduced this approach. Also, it works on condition. That means when data in a high-intensity space needs to mark data in a low-size space. Or, the detail variation within the space below should be higher. The first step is to make a standard one, here are the common features for making zero mean and unit variance for each feature. The second step is to remove the main part. It includes the following steps:

- Create a data covariance matrix.
- Calculate the eigenvectors of this matrix.

We use Eigen vectors corresponding to the largest eigenvalues. That is to rebuild a large part of the real data difference. Therefore, we are left with a small number of eigenvectors. And there may be some data loss in the process. However, the most important difference should be maintained by the remaining eigenvectors sources. The main component is called the high exchange direction represented by the eigenvector co-variance matrix and its corresponding eigenvalues represent magnitude.

VI. LINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis (LDA) is a supervised learning method. It finds space for a new data prediction feature with high classification between classes. Calculate between class and middle class to spread metrics. Group spread means a small difference within each category and between a group break means a significant distinction between the definitions of each group category. It includes the following steps:

- Make the D-face data in the right place.
- Each class counts d-dimensions meaning vectors
- Build between a scatter class matrix and within a scatter- class matrix.
- Calculate the eigenvector of the matrix
- select k gene-vectors corresponding to k-size gene values to form (d, k) conversion matrix W
- Project X (input data) on matrix W

VII. CLASSIFICATION

Classification is a data mining technique in which label the known facts in order to know the unknown facts. It can be used for the prediction task. Decision tree, rule based, Neural network , probabilistic approach etc: are some of the examples of classification algorithms. Separation is a two-step process, one step learning and the other step separation. In the learning step where the separation model is built and in the partition phase where the model is used to predict a class data label. The two purposes of classification model are descriptive modeling and predictive modeling. In descriptive modeling we have describing about data and in predictive modeling we are predicting the future pattern. General approach to solve a classification problem is to identify the test model and after that build models with good generalization capability.

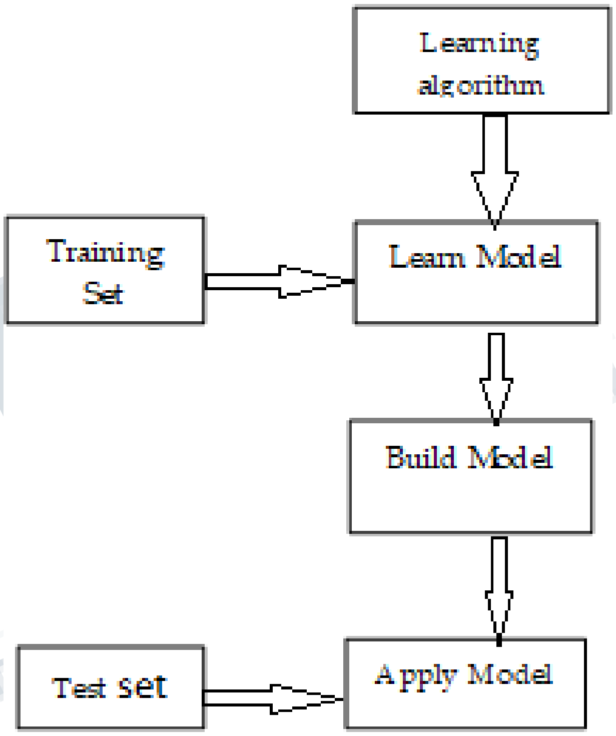


Fig. 3. Classification

VIII. COMPARISON STUDY OF THE EXISTING SYSTEM

Predictive models are important in personalized medicine because they help to identify high-risk individuals early on, based on known epidemiological and clinical risk factors. Vector Machines support models, K's closest neighbor, Nave Bayes, and Decision Tree classification models used to train models for categorizing the two forms of breast cancer using features selected at different threshold levels. Several empirical research have used machine learning and soft computing techniques to treat breast cancer. Performance can be measured on sensitivity, specificity, precision, accuracy, and the ROC curves. Feature selection is a necessity for model construction in machine learning since it involves selecting a subset of relevant attributes from a large number of potential subsets. In order to build an effective predictive model, feature selection is critical.

In a segmentation and verification of lung nodule image identification using ANN, pre-processing and augmenting the picture regions obtained from the segmentation phase are the first steps in the classification process. The expanded dataset is then used to train an ANN using back propagation. Finally, a neighboring image classification and rejection algorithm is used to limit overlapping detections.

Performance of classification model can be evaluated using confusion matrix. It gives the counts of correctly and incorrectly predicted test records and displays the result in table format. Performance was tested for accuracy and error. Accuracy has been assessed by dividing the correct predictive value by the total number of predictions. The error level was checked by dividing the negative prediction number by the total number of predictions

A. Support vector machine

Support Vector Machine is one of the simplest algorithms that can provide remarkable accuracy with minimal computational power. It can be used for reversal and separation functions. he purpose of this algorithm is to detect a hyper plane in the N space that clearly separates the data points. It is possible to find multiple planes, which is more suitable for high-resolution, which means that there is a line with a higher distance between the data points of both classes. This hyper plane separates the two classes that help to separate the data points.

The supporting vectors of data points are close to the hyper plane and influence the position and position of the hyper plane. These support vectors plays a key role in the maximization of separation between two classes. Changes in these points will affect the position of the hyper plane. These all are the crucial points which needs to consider while building our SVM.

B. K-nearest neighbor

The closest neighbor to K (KNN) is one of the monitored algorithms that can be used for regression and classification. The principle behind KNN is that similar things are likely to close each other. So based on the distance between the data points, we are measuring the similarity between data points. One of the main challenges in KNN is to choose the best value of K where we can classify the given data points efficiently. One of its main disadvantages is slower as the size of data increases.

C. Naive Bayes

The Naive Bayes is a way of dividing according to the Bayes theory. The Naive Bayes divider thinks that the existence of a certain element is not related to the existence

of any other element. It's a way of differentiating according to the Bayes 'Theorem and the idea of independent autonomy. It is able to deal with continuous and unique data. It can handle a large number of predictors and data points. It is fast and can be used to make real-time predictions. It is able to deal with continuous and unique data. It can handle a large number of predictors and data points. It is fast and can be used to make real-time forecasts. It is unaffected by non-essential characteristics. It is unaffected by non-essential characteristics. These are most commonly used in sentiment analysis, filtering spam messages etc. Due to its naive nature its easy and fast method to implement.

D. Decision Tree

This is one of the powerful tool or algorithm which explicitly represents decisions and helps in decision making. This can be used in both classification and regression. It is used to create data models that predict class labels or values to aid in decision-making. Decision Trees are a type of Supervised Machine Learning in which data is continually separated based on a parameter. The models are created using the training data that has been given into the system. The pruning tree is a tree-like flowchart, in which each internal node defines a test in the attribute, each branch represents a test result, and each leaf node (terminal node) holds a category label.

E. Artificial neural network

ANNs are another common computerized system that is programmed and based on the functioning of neurons in the human brain. ANNs offer ways to manage many different parameters where a specific analytical model is missing or difficult to build. ANNs also provide a consistent way of looking at large amounts of data as well as simple ways to assess the potential outcome of a complex problem with a set of specific conditions. Artificial neural networks are known for flexibility, which means they are flexible as they learn in the first training and the next run provides more details about the world.

IX. CONCLUSION

The main objective is to study the impact of feature selection in hybrid classification of high dimensional data sets. In this paper we are focusing on various feature selection methods and its impacts on the accuracy of the system and a comparative study on various classification algorithms on the selected features. Curse of dimensionality is a major challenge and will reduce the accuracy of the system. Here comes the importance of a suitable feature selection. Predictive models are important in personalized medicine because they help to identify high-risk individuals early on, based on known epidemiological and clinical risk factors. In the medical field prediction system should be accurate and effective and has to be affordable by the common people. Different classification approaches exists in the healthcare analysis. Many challenges are there while applying these algorithms. One of the main challenges is data redundancy, which can affect the accuracy of our entire system. In order to overcome this barrier we need to have some feature selection approaches. Here we are looking forward to use the machine learning aspects in the field of Oncology department more specifically in the area of Breast Cancer. The most significant fields of research are analytical identification of crucial diseases such as breast cancer. The WDBC UCI dataset, which has 569 instances and 32 variable attributes, is the most often used data set for detecting breast cancer. The information offered by the UCI repository is extremely useful in identifying the characteristics that are important in determining the type of breast cancer a person has.A good selection of features can be key to creating a well-balanced and accurate separation. In this paper we have learned various ways to select features and then applied classification algorithms on the selected features and performed a comparison study about the accuracy of the system.

REFERENCES

[1] F. Padillo, J.M. Luna†, and S. Ventura, "An evolutionary algorithm for mining rare association rules: a Big Data approach"Department of Computer Science and Numerical Analysis 978-1-5090-4601-0/17 c2017 IEEE

[2] A. Strehl and J. Ghosh "Cluster ensembles – A knowledge reuse frame- work for combining multiple partitions," Journal of Machine Learning Research, pp.583-617, Feb. 2002.

[3] A. Fred and A. K. Jain. "Combining Multiple Clusterings Using Evi- dence Accumulation," Analysis, vol. 27, no. 6, pp. 835-850, 2005.

[4] Bichen Zheng, Sang Won Yoon, Sarah s Lam,"Breast Cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms, " Expett Systems with Applications 2013

[5] Smitha, Hareesh, Seema, Rajesh "A Neural Network Based Breast Cancer Prognosis Model with PCA Processed Features 2016 Intl. Con- ference on Advances in Computing, Communications and Informatics (ICACCI), Sept. 21-24, 2016, Jaipur

[6] Salama,Abdelhalim, Zeid "Experimental Comparison of Classifiers for Breast Cancer Diagnosis " 978-1-4673-2961-3 2012 IEEE

[7] Sadegh, Mohammad "Application of K-Nearest Neighbor (KNN) Ap- proach for Predicting Economic Events: Theoretical Background " S B Imandoust et al. Int. Journal of Engineering Research and Applications Vol. 3, Issue 5, Sep-Oct 2013, pp.605-610

[8] Onan " A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer". Expert Syst Appl 42(20):6844–6852 (2015)

[9] Addeh A, Demirel H, Zarbakhsh P " Early detection of breast cancer us- ing optimized ANFIS and features selection." In: 2017 9th international conference on computational intelligence and communication networks (CICN), Girne, 2017

[10] ., ZhigangZeng." A new automatic mass detection method for breast cancer with false positive reduction", Neurocomputing 152 , 388–402. 2015

[11] B. C. Majoor, A. M. Boyce, J. V. Bove´e, V. T. Smit, M. T. Collins, A. M. Cleton-Jansen, et al., "Increased risk of breast cancer at a young age in women with fibrous dysplasia," Journal of Bone and Mineral Research, vol. 33, pp. 84-90, 2018.

[12] C. Kullberg, J. Selander, M. Albin, S. Borgquist, J. Manjer, and P. Gustavsson, "Female white-collar workers remain at higher risk of breast cancer after adjustments for individual risk factors related to reproduction and lifestyle," Occup Environ Med, pp. oemed-2016- 104043, 2017.

[13] S.Sandhiya. Y.Kalpana, "A Segmentation and verification of Lung nod- ule image identification using ANN," International Journal of Computer Science and Information Security (IJCSIS), Vol. 16, No. 11, November 2018