



ISSN: 0976-3031

Available Online at <http://www.recentscientific.com>

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 11, Issue, 11 (A), pp. 40000-40007, November, 2020

**International Journal of
Recent Scientific
Research**

DOI: 10.24327/IJRSR

Research Article

FEATURE SELECTION METHODS: A SURVEY FOR EFFECTIVE DATA CLASSIFICATION

**Sivakumar Venkataraman¹, Subitha Sivakumar², Saranya K.R³,
Appavu alias Balamurugan⁴ and S.Sasikala⁵**

¹Department of Health Information Management, Botho University, Gaborone, Botswana

²New Era College of Arts, Science and Technology, Gaborone, Botswana

³Dept. of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, Tamilnadu, India

⁴Dept. of Computer Science, Central University of Tamilnadu, Thiruvavur, India

⁵Dept. of Computer Science and Engineering, Velammal College of Engineering and Technology, Madurai, India

DOI: <http://dx.doi.org/10.24327/ijrsr.2020.1111.5605>

ARTICLE INFO

Article History:

Received 15th August, 2020

Received in revised form 7th September, 2020

Accepted 13th October, 2020

Published online 28th November, 2020

ABSTRACT

Data is growing every day. This daily growing data is called Big Data. The capturing and accessing and preserving of these data is different and difficult. If capturing and preserving of data became very difficult, then classifying data according to the need is also a difficult process. These modern days datasets are high dimensional that is with hundreds of features and keeps on increasing in future. But these hundreds of data are not required for an effective data classification. The aim of this survey is to give a complete review on the different feature selection methods to improve the accuracy of the data classification.

Key Words:

Dimensionality reduction, Feature Extraction, Feature Selection and Data Classification

Copyright © Sivakumar Venkataraman et al, 2020, this is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

A vast number of datasets is available throughout the world in the Information and Communication Technology industries. This type of datasets needs to be transferred to a useful dataset, by extracting the relevant data from the dataset. Apart from extraction needs to include the added procedures like pre-processing used to clean the data, combining the data, transfer the data, pattern assessment and demonstration. After completing the above-mentioned processes, we can be able to use the extracted information in various applications.

Data classification is a process of allotting or categorizing a data to a particular class or group. For each business requirements different kind of data are needed, so we need to classify according to that. On a basic level, classification process makes the data easier to locate and retrieve. Big data are high prone of having duplicate data. Hence the very important point while classifying the data is we need to

eliminate the duplicate data which in turn prevent from using more memory and also increases the speed of search. It is merely a data management process. The classified data can be represented as public or private. Therefore, publicly available data has less sensitive information with less security requirements, where the restricted data has highly sensitive information with high security. This form of data This sort of information characterization is regularly the beginning stage for several creators, trailed with further identifications and use the tag methods to labeling the relevant data to the creators, quality and other classification. The classification task predicts or identifies the label of an unlabeled data using a classification model which is known as classifier. The classification accuracy is mainly failing on the training data where used in the learning and developing the classifiers.

The challenging part in dealing with high dimensional data is to build the classifier, since it contains irrelevant and duplicate data as mentioned earlier. The irrelevant features and the

*Corresponding author: **Sivakumar Venkataraman**

Department of Health Information Management, Botho University, Gaborone, Botswana

redundant features will take to incorrect classifications. So, to obtain the most relevant features the feature selector methods are used (Mehnaz & Quadri, 2013). Feature selection is a method which reduces the input data used to process and find the optimum features. In this paper, first the widely used feature selectors are analyzed, second the dimensionality reduction and the methods used for reduction are discussed, third the feature extraction methods are briefly described, fourth the feature selection procedures are explained and finally applications of data classification is presented followed by conclusion in section 6.

Dimensionality Reduction

Dimensionality reduction used to reduce features from the dataset by removing the unwanted features which are not necessary for any particular task (Lakshmi *et al*, 2014). Hence removing the unnecessary variables from the data are called as dimension reduction. Data which requires two to three dimensions to represent is called as high dimensional data. Large amount of data are created daily. Because of increasing in data it has become a challenge in handling the data and also in imagining the data. By using this method, it is easy to learn by using the email classification problematic, by classifying whether the email is spam. Sometimes, the selected features might overlap. In other scenario, the classification falls with humidity feature and rain fall feature, sometimes it can failure due to each feature are correlated with each other to a high degree.

However, some of these features may overlap. In another condition, a classification problem that relies on both humidity and rainfall can be collapsed into just one underlying feature, since both of the aforementioned are correlated to a high degree. Though feature selection and dimensionality reductions both aims at reducing the number of features, but both have some difference. The main difference is the extracted features will be the subset of the main features (Lakshmi *et al*, 2014). Feature selection mainly applied on the attributes and it eliminates the attribute according to the value of the variance. But dimensionality reduction concentrates on creating new dimension depending on the value of covariance. Dimension reduction are focused on the feature selection and extraction.

$$\begin{aligned}
 & a+b+c+d = e \\
 \Rightarrow & a+b = ab \\
 \Rightarrow & ab+c+d = e \quad (\text{extraction}) \\
 \Rightarrow & c = 0 \\
 \Rightarrow & ab+d = e \quad (\text{selection})
 \end{aligned}$$

Feature Extraction

Feature extraction used to reduce the dimensionality through which the original data is reduced for further processing. This method combines the features by reducing the size of the dataset without distributing the quality of the original datasets. A trait of these huge informational collections is countless factors that require a great deal which requires more additional resources for processing the data.

The importance of this method is to reduce the size of features without losing the essence of the original data. By using this method, the redundant features can also be reduced. When the size of the dataset is reduced, the execution speed will be increased in the classification. There are different feature

extraction methods which reduces the size of the dataset without compromising the quality of the dataset. The common feature extraction techniques include:

Missing Value

A missing values can connote various things in the data. Maybe the data was not accessible or not pertinent or the function didn't occur. It may be the case that the entered the data didn't have the foggiest idea about the correct worth, or missed filling in. Data mining strategies shift in the manner to treat missing qualities. In simple, missing value appears as there is no value stored for that feature.

Missing value plays very important task in data classification. The missing value occurred to various reasons like human error, corrupt data, duplicate data, null data etc., But there are different ways to handle missing values in a data set. Looking into the missing values is most significant in the data-pre-processing stage because during the machine learning process most of the methods would not on how to handle the missing values.

Table 1 Missing values

ID	TEST	T1	T2	T3
1	Patient 1		0.021	0.221
2	Patient 2	0.925	1.376	0.435
3	Patient 3	1.034	2.675	1.982
4	Patient 4	1.406		1.823
5	Patient 5	2.591	0.491	4.376
6	Patient 6	4.44	0.222	3.946
7	Patient 7		3.856	0.29
8	Patient 8	2.465	0.321	1.678
9	Patient 9	0.245		2.782
10	Patient 10	1.385	3.652	0.481

Deleting rows

Deleting rows are used to deal with the empty data. If a row has a null value, we need to remove that particular row and if a column has more number of missing value then that particular column has to be deleted. Deleting rows can be implement where the dataset are more with sufficient samples.

Replacing with Mean/Median/Mode

This method can be applied on a feature which contains numerical data such as the age of a person or a ticker fare. Since this method deals with numerical data it is easy in finding the mean, median and mode of the features. After finding it should be replaced by the missing values. This method is statistical process to handle the missing values in the dataset when the data's are linear.

The mean can be obtained by adding the values in the dataset and by dividing the total using the number of values. The mean can be obtained by looking in to the middle value when the numbers are arranged in the ascending order. The mode is the number that occurs most often in the dataset.

Assign unique category

Features that have defined values of data like age, department, martial status etc., are consider as a category feature. If there are any missing values, it can be assigned by another class, example 'U' as unknown class. By this, the datasets will be added with more informations where it leads few variance in the results.

Predicting the missing values

By using the machine learning methods, the missing values can be identified and to be removed and will have good accuracy in the result. If the missing value is having big difference in the variances, it may affect the result. The common method in handling the missing value is to eliminate the record from the dataset before the analysis part beings.

Decision Tree

A decision tree is the supervised machine learning algorithms which can be implemented for the regressions and classifications problem. Building up a decision tree can be done by identifying the features which returns the maximum information gain value.

A decision tree is mostly applied for the operational research, such as analysing, where to helping to find a approach to achieve the upmost goal and the decision tree is well known algorithm in machine learning.

A decision tree looks like a tree in structure and contains internal node which is called as 'test', each division are the result of the node and each leaf node are the class labels. A decision tree looks like a flow chart where it can be used in determining the flow of the features arranged in the dataset. The decision tree uses the supervised learning method to find the results. This method has more accuracy. Decision tree model in machine learning used to create: Induction and Pruning.

Induction is the place the decision tree is build. Pruning removes the unwanted structure in the tree where to make the tree easy to interpret.

Low Variance

Variance refers on how the information's are spread. If the variance shows zero then the data's are identical. The non zero variances are positive. The data with less variance indicates the data values are near to the mean and to others. If the variance is high then the data value are far from the mean and from other. Variance refers to the avg.sqaure distance to each mean point. To find the variance,

Step 1: to calculate the mean for the dataset.

Step 2: to find the distance value, difference between the data value and mean is calculated

Step 3: Squaring the distance value

Step 4: Adding the squaring values and dividing with the total data number

After finding out the variance, feature with high variance are removed from the data set.

High Correlation

Data columns with very similar nature will have similar data. In that case, those similar data might not help out in feature extraction. Removing those data would not decrease more on the available size of information (Malavika & Selvam, 2015). If needed to remove any of the decidedly correlated values, before that necessity to find the correction among the pairs through linear correlation node. This finds the coefficient of each pairs in the datasets, as the Pearson's Product Moment Coefficient

(Malavika & Selvam, 2015; Chelvan & Perumal, 2017) and to each pairs as Pearson's Chi Square Value (Chelvan & Perumal, 2017).

Pearson's Product Moment Coefficient

The correlation Coefficient shows the pathway and linear relation among the features and denoted as Pearson's Product Moment Coefficient, this can have the values within the range of -1 through 0 to +1. The sign (+ or -) of the correlation affects its interpretation.

If the sign is (+) for the correlation, then the value for the one feature will increase and will increase for the other. (ie) lets consider Average, if the correlation for Average is (+), then the value for height will increase and will do the same for the weight. If the sign is (-) for the correlation, then the value for the one feature will increase and will decrease for the other. (ie) lets consider Average, if the correlation for Average is (+), then the value for height will increase and will decrease the weight. When the variables are not related to each other, if one variable is increased, then there may be decrease or increase on the other, this is the zero correlation.

The correlation coefficient is calculated by using the

$$r = \frac{\sum xy}{NS_x S_y}$$

Here; x, y are represented as the deviation values as

$$x = X - \bar{X} \text{ and } y = Y - \bar{Y}$$

$S_x, S_y \Rightarrow$ denoted as the sample standard deviations,

$$S_x = \sqrt{\frac{(X - \bar{X})^2}{N}}$$

Thus, the correlation is average cross-product consistent which is divided by the standard deviation.

Pearson's Chi Square

If the column has nominal data, then the correlation is found using Pearson's Chi Square method is used. The Chi-Squared test is a statistical process used to identify the frequency for the definite features which match the predicted frequency. The chi-square is calculated by

$$\chi^2 = \sum_{i=0}^k \frac{(O_i - E_i)^2}{E_i}$$

By O as Observed value, E as expected Value and i referred as the location in the table.

If the chi-square value is low, the dataset is having a high correlation. If the chi-square value is high, then the dataset is having a less correction. If the chi-square value is zero, the both the observed and expected values are same.

Principal Component Analysis (PCA)

PCA algorithm refers to the feature extraction technique which uses the orthogonal linear projection method to find the original variances of the features. This method can be used in reducing the size of the dataset from large to small dataset where in not compromising the information's in the original dataset.

The actual coordinates for the dataset is converted to the new coordinates in the principal component. By this the size of the features can be reduced through selecting the optimal features (Gerard, 2012). In PCA there are five stages; to find the covariance matrixes, to find the Eigen value and Eigen Vector, then by using the value and vector the optimal features vector are selected and transmute the data's to the selected vector to reduce the size of the dataset (Paul *et al*, 2013; Paul *et al*, 2013). Stages are;

Stage 1: Data Normalization

Data normalization is the fundamental step in data mining and also known as data preprocessing. In the first stage, the data is normalized where the PCA can work accurately. This normalization can be performed through subtract the mean value from the numbers in the equivalent columns. That is nothing but transfer the feature from data format to another data format which helps the dataset to process effectively. Then, to remove or minimize the dataset by removing the duplicate features which also helps to the dataset to perform better. In simple, data normalization helps to organize the features in the dataset to maximize the cohesion of the entities type.

Stage 2: to find the covariance matrix

When the data set is having two dimensional, the covariance matrix will be 2X2.

$$\text{Matrix (covariance)} = \begin{bmatrix} \text{Var}(x) & \text{Cov}(x, y) \\ \text{Cov}(x, y) & \text{Var}(y) \end{bmatrix}$$

$$\text{Var}(x) = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

Stage 3: to find the Eigen values and vectors

In the stage 3, the Eigen Values and Vectors are calculated. λ is the eigen value for a matrix A.

$$\det(\lambda I - A) = 0$$

"I" represents identity matrix of the same dimension as A which is a required condition for the matrix subtraction and the 'det' is the determinant of the matrix. For each eigen value λ , corresponding eigen vector v, can be found by solving:
 $(\lambda I - A)V = 0$

Step 4: selecting the feature and to form a feature vectors.

Eigen values are arranged in descending order where to have the proper the components as in connotation list. If the data set has a number of n features, the Eigen values and Eigen vectors will be n numbers. That means, the uppermost value of the eigen vector is the principal component for the dataset. If the dataset demission to be reduced, then select the first p eigen value and leave the other values. The Feature vector can be formed by

$$\text{Feature Vector} = (\text{eig}_1, \text{eig}_2)$$

Step 5: To form the Principal Components

In this stage the principal components are formed by transmuting the feature vector with the original datasets (Paul *et al*, 2013)

$$\text{New Data} = \text{Feature Vector}^T \times \text{Scaled Data}^T$$

PCA can be applied to different real time applications to select the relevant features like prediction of warning level in aircraft accidents (Arockia & Balamurugan, 2013).

Random Forest / Ensemble Trees

Random Forrest also refers to supervised learning algorithms and creates multiple decision trees, then links the tree to obtain the utmost accurate prediction. This random forest is trained by 'Bagging' approach. The main objective of this bagging approach combines the learning model to improve the results. This algorithm can be implemented for classifications and regressions.

The decision tree created by the random forest will have the topmost as the hyper parameters. This is an flexible algorithm were to use in the machine learning. Random Forest obtains the best output at most in times. This algorithm is trained by using the "Bagging" process (i.e) to combine the learning models to increase the result.

In simple, by using the random forest algorithm the multiple decision trees are built where to have more accurate and more stable prediction.

This hyperparameters created can increase the prediction and it can also speed up the model execution time.

The hyper parameter (ie) the n estimator is the number of the multiple decision tree created on before the maximum voting is included. Higher the number of trees higher the performance and makes the predictions more stable, but at the same time it slows down the computation.

Max-features is another hyperparameter, where the maximum number of features are used to split a node. Finally, min_sample_leaf is parameter which refers the minimum number of leafs are used to split the internal node.

Feature selection and its methods

Feature selection is used in reducing the features in dataset or finding the most relevant inputs. In simple, feature selection is used to reduce the input values for processing and analysing or in finding the optimal features. The Big Data has vast features, which is not required in building the models. Example, customer dataset might have 500 features which describes the features of the customer, but all the features may not be used to build the model where the other features might not be related, be redundant or be benefitable.

In data classification, the feature selection plays an important role in selecting the relevant features in building the models. Because of selecting the optimal features, it also increases the performance of the model. Sometimes, when using the unwanted features in building the models, it holds more memory. Hence, it reduces the quality and obtaining the suitable pattern will be difficulty. When the datasets are huge volume in size, the datamining techniques will need large number of data to be trained by using the training datasets. Thus, the feature selection method is used in solving the two main issues; more features of having less values, less features having high values. Datamining has several methods and procedures which can be used in obtaining the best features where to help the models to do better and efficient.

Filter Method

Filter method are mostly used as in the preprocessing step. The filter method uses the statistical approach to select the number features independently by using the machine learning algorithms, that is based on the numbers assigned by the multiple statistical approach the features are selected on the correlations of the output features (Mwadulo, 2016; Arul & Elavarasan, 2014).

The feature selection approaches can be used to reduce the number of inputs where those inputs are most relevant features in building the model and also help to predict the target feature. There are various methods to find the correlation coefficient like Pearson Correlation, Linear Discriminant Analysis (LDA), ANOVA i.e. analysis of variance and Chi-Square.

Linear Discriminant Analysis (LDA)

LDA method are mainly applied to reduce the dimension of the dataset and in obtaining the useful pattern classifiers. The main aim of LDA is to implement dataset on a lower dimensional space which prevents over fitting and reduces the computational cost. This analysis method is same as the Principal Component Analysis; addition to that it finds the component axes which maximize the variance of our data (Zhao et al, 2019).

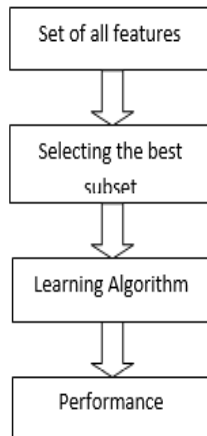


Figure 1 Filter Approach in Feature Selection

Given the data matrix $X = [x_1, \dots, x_n] \in \mathbb{R}^{d \times n}$, here d , n represents the dimension of the actual data size and sample numbers. The main aim of this model is to train the linear transformation matrices $W \in \mathbb{R}^{d \times m}$ ($m \leq d$) and helps in mapping the vast size of dataset $x \in \mathbb{R}^d$ to small size of dataset $y \in \mathbb{R}^m$ by

$$Y = WX$$

Where X combined as $x = [\pi_1, \pi_2, \dots, \pi_c]$, c represents the class numbers, $\pi_i \in \mathbb{R}^{d \times n_i}$ represents the dataset in class i , n_i represents sample numbers of data in class i . To find the distinguishing features, the best prediction matrix can be attained by (Zhao et al, 2019),

$$\min \text{Tr}(WTStW) - 1(WtS_wW)$$

where $\text{Tr}(\cdot)$ represents the tracing of matrix's, where the class scatter matrixes are defined as

$$S_w = \sum_{i=1}^c \sum_{x \in \pi_i} (x - \bar{x}_i)(x - \bar{x}_i)^T$$

where $\bar{x}_i = \frac{1}{n_i} \sum_{x_j \in \pi_i} x_j$ represents the sample mean in class i and the total class scatter matrix is

$$S_t = \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ represents the mean for all samples.

The solution of the objective function can be reduced to solving the following eigen values decomposition problem:

$$S_t^{-1} S_w W = W \Lambda$$

where Λ is the eigen value matrix of $S_t^{-1} S_w$. Finally, the transformation matrix W are comprised with eigen vector of $S_t^{-1} S_w$ matching with first m lowest eigen values.

Relief

The RELIEF method (Sanchez et al, 2007) is used to find the features quality in the dataset. From the randomly selected features $X_i = \{x_{1i}, x_{2i}, \dots, x_{mi}\}$, the RELIEF algorithm searches the adjacent neighbors features: first feature within the class, referred as H where H is nearest Hit and the second feature from the other classes, referred as M where M is the nearest miss. An extension, ReliefF algorithm helps in dealing with incomplete and noisy data. It also deals with multi-class problem.

The Relief algorithm families are widely used because they can be implemented in all cases, like interacting among the features and may capture local dependencies that other methods miss (Sanchez et al, 2007). Comparatively Relief feature selection method selects good features than the other methods and it also reduces the usage of huge memory space and the error rate is also minimized (Aljumah et al, 2013).

Relief Algorithm

Input: for each training instance a vector of attribute values and the class value.

Output: the vector W of estimations of the qualities of attributes.

set all weights $W[A] := 0.0$;

for $i := 1$ to m do begin

randomly select an instance R_i ;

find nearest hit H and nearest miss M ;

for $A := 1$ to a do

$W[A] := W[A] - \text{diff}(A, R_i, H)/m + \text{diff}(A, R_i, M)/m$;

end;

ReliefF Algorithm

Input: for each training instance a vector of attribute values and the class value

Output: the vector W of estimations of the qualities of attributes

set all weights $W[A] := 0.0$;

for $i := 1$ to m do begin

randomly select an instance R_i ;

find k nearest hits H_j ;

for each class $C \neq \text{class}(R_i)$ do

from class C find k nearest misses $M_j(C)$;

for A:= 1 to a do
 $W[A] := W[A] - \sum_{j=1}^k \text{diff}(A, R_i, H_j) / (m.k) +$
 $\sum_{C \neq \text{class}(R_i)} \left[\frac{P(C)}{1 - P(\text{class}(R_i))} \sum_{j=1}^k \text{diff}(A, R_i, M_j(C)) \right] / m.k;$
end;

The ReliefF Algorithms are used to deal with unfinished, useless features. Similar to Relief algorithm, the ReliefF algorithm picks an instances R_i , then searches k that is the nearest neighbor in same class as H_j and k nearest neighbor from other classes as $M_j(C)$. Also finds the $W[A]$ for all attributes A as depends on R_i , H_j and $M_j(C)$. In the new ReliefF formula is very same as the Relief, but except the average contributions of H_j and $M_j(C)$. This $M_j(C)$ contribution in each class are weighted with the prior probability of that class $P(C)$.

Correlation-based Feature Selection (CFS)

CFS are common filter method used to rank the subset features on the correlations heuristic evaluation function (Sanchez *et al*, 2007). This method focused on the subset which has the features. This feature are mostly correlated within the classes and the very correlated among each other. Here, the unrelated data must be eliminated, because the features are having less correlation within the classes. The redundant data must be analyzed thoroughly because the features might be high correlation with other lasting features. The features will be selected on the performance in predicting the class. The correlation based feature selection feature subsets evaluation function is

$$M_S = \frac{k \bar{r}_{cf}}{\sqrt{k + k(k-1) \bar{r}_{ff}}}$$

Here M_S represents the heuristic subset feature S having k features, $\bar{r}_{cf}$ represents the mean feature- class correlation and $\bar{r}_{ff}$ represents the average feature- feature inter correlation.

Fast Correlation- based filter (FCBF)

FCBF is a Symmetrical Uncertainty (SU) where defines the ratio among the InformationGain (IG) and Entropy (H) of the feature x and feature y (Sanchez *et al*, 2007)

$$SU(x, y) = 2 \frac{IG(x/y)}{(H(x)) + (H(y))}$$

IG are calculated by

$$IG(x/y) = H(x) + H(y) - H(x, y)$$

$H(x)$, $H(x, y)$ represents the entropy and combined entropies. These handles the vast size of dataset and is most effective in handling the noisy and redundant data.

Interact

Interact are more likely to FCBF filter since it makes use of the symmetrical uncertainty (SU) (Sanchez *et al*, 2007). This comprises the consistency contribution (C-Contribution), were indicates on how meaningfully the removal of each features will disturb the consistencies. It has two main stages. First Stage based on the Symmetrical Uncertainty value the features are ranked in descending order. Second Stage, the evaluation is done from the lass feature from the list. When the consistency contribution value for a feature is less compared to the

threshold value, then the feature is eliminated. If the value is not less, then the feature is selected. By using the Interact algorithm, this can handling the feature interactions and also it selects the most optimal features.

Wrapper Methods

Wrapper approach is used to train a model by using a subset of features. This is not like other methods, the features can be included or removed from the subset. Wrapper Method is high expensive. Few of this method are forward feature selection, backward feature elimination and recursive feature elimination (Mwadulo, 2016; Arul & Elavarasan, 2014).

This method are divided as Sequential Selection Algorithm and Heuristic Search Algorithm. This algorithm performs the frontward or reverse selection (Nancy & Balamurugan, 2013) by using the Sequential Forward Selector method. First it starts with the null sets and includes one feature, then obtains the maximum value for the objective functions. Secondly, adds the rest features one by one and creates the new subsets. The features with highest classification accuracies are added in the subset permanently.

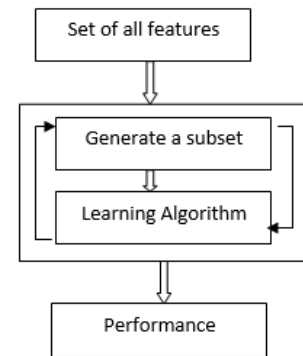


Figure 2 Wrapper Approach in Feature Selection

This approach is done repeatedly till the essential features are included in the subset. this is an naïve Sequential Forward Selection method were the dependency among the features are not considered. This method is not feasible for feature subsets selection for a high size of features with high dimensional features. SFFS are flexible when compared to the SFS algorithm, since this method adds the additional backtracking stage. SEFS and SFS has the first step as the same where it includes one feature at a time. SEFS has a step which eliminates on feature from the subset at a time and creates a new subset. Because of this, the value of the objective function is increased, so that feature should not be considered and takes to the first step for new subset, until the essential features are obtained. Here the nested subsets are obtained by SFS and SFFS. The forward inclusion was unconditional, when the two highest correlation features are added then the maximum performance is obtained by using the SFS.

This SFFS was used to avoid nesting effect. The parameter r used in the ASFFS represents the number of features to be included. The o parameter represents the number of features to be removed. This ASFFS method obtains the minimum redundant subset features. Another method is Sequential Backward Selection (SBS), same as SFS. SBS begins from the completed subset and starts to eliminate one feature where the elimination produces the less effect in the prediction performances.

Heuristic Search Algorithm

By using Heuristic Search Algorithm, the various subset are generated by two methods; first method to search by using the search space and the second method to generate problems. GA method uses the Darwinian principle of 'survival of the fittest'. The fitter chromosomes have higher probability to be kept in the next generation or selected into the recombination pool using the tournament selection methods. If the fittest individual or chromosome in a population cannot meet the requirement, successive populations will be reproduced to provide more alternate solutions.

The crossover and mutation functions are the main operators that randomly transform the chromosomes and finally impact their fitness value. The evolution will not stop until acceptable results are obtained. GA can work efficiently with large search spaces. GA can help in finding feature subset selection because they are effective for fast global search of large, non-linear spaces. Thus, unlike most standard heuristics, GA uses information of a population (individuals) of solutions when they search for better solutions.

Embedded Method

Embedded method combines the filter and wrapper method (Arul & Elavarasan, 2014). Embedded method uses the algorithm has the own built in feature selection methods like LASSO and RIDGE regressions.

In embedded approaches, the feature selection algorithms are integrated along with the learning algorithms. The filter methods finds the relation between the features through the correlation by using the dependent features. The wrapper methods finds the practicality of the subset features through training the models. When comparing the both methods, the filter method is more faster than the wrapper method because the filter method include in training the models and also wrapper method is more expensive. Filter methods uses the statistical procedures in evaluating the subset features where the wrapper uses the cross validation approach.

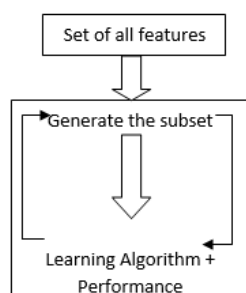


Figure 3 Embedded Approach in Feature Selection

Thus wrapper methods are essentially solving the real problem but it is computationally more expensive compared to filter methods due to the repeated learning steps and cross validation. The Embedded methods are similar to the wrapper methods because it also used to optimize the objective function or performance of the learning algorithm or model.

Applications of data classification

Data classification methods can be implemented different kind of applications like in marketing field, client relations

managements, engineer fields, medical fields, proficient prediction, internet fields and mobile fields. The datamining classifiers used on healthcare like diabetes or the prevalence of diabetes increases in patient, the outcome shows in stopping the smoking decreases the rate of diabetes. As to (Aljumah *et al*, 2013) states smoking is a main reason which also increase the diabetic ratio. The feature selection can improve the biased aptitude in the medical dataset called MultiFiltration Feature Selection (MFFS), which includes feature extraction, feature subset selection, feature ranking and classification (Balamurugan *et al*, 2016). Shapely Value Embedded Genetic Algorithm (SVEGA) are depended on Genetic Algorithm, were the features are added and removed. This method helps diagnose the Breast cancer by reducing the dimensionality and to improve the accuracy (Balamurugan *et al*, 2015). Similarly, the main problem in modern finance is identifying the efficient way to summarize and visualize the stock market data. Data mining techniques have been applied to reveal the hidden patterns in the data and to predict future scope and behaviors in financial markets (Radhakrishnan *et al*, 2013).

In business world, Customer Relationship Management (CRM) is very important. CRM gives us a customer classification and prediction, which is used for the improvement of the business process (Mangal & Jain, 2016). Growing popularity of the World Wide Web, many websites typically experience Lac of visitor's every day. Data mining techniques like association rules, classification and clustering are considered very useful in web usage mining. Data mining is gaining more popularity because of its power to extract knowledge from voluminous data where it is beyond the reach of traditional techniques of knowledge discovery and human comprehension.

CONCLUSION

In conclusion, this paper shows the study to find out the efficient method of feature selection for effective classification of the data. The application works carried out by different researchers were discussed. Thus data classification is a diverse process that involves various methods and criteria for sorting data within a database or repository and moreover selecting the good features which improves the classification accuracy.

Acknowledgements

Authors acknowledge their colleges, co-authors and the university supported us to motivate and publish the research work.

References

1. Abdullah A.Aljumah., Mohammed, Khubeb Siddiqui and Mohammed Gulam, Ahamad. (2013). Application of Classification based Data Mining Technique in Diabetes Care. *Journal of Applied Science*, 13(3) 416-422.
2. Amarnath, B & Appavu alias, Balamurugan. (2015). A Novel Hybrid Feature Selection via Information Gain based on Frequent Pattern Analysis. *International Journal of Conceptions on Computing and Information Technology*, 3(2), 2345- 9808.
3. Appavu alias, Balamurugan, Saskala, S & Geetha, S. (2015). A Novel Feature Selection Technique for Improved Survivability Diagnosis of Breast Cancer.

- International Symposium on Big Data and Cloud Computing, 16- 23.
4. Appavu alias, Balamurugan., Sasikala, S & Geetha, S. (2016). Multi Filtration Feature Selection (MFFS) to Improve the Discriminatory ability in the clinical data set. *Applied Computing and Informatics*, 12, 117-127.
 5. Arockia Christopher, A.B and Appavu alias, Balamurugan. (2013). Feature Selection Techniques for Prediction of Warning Level in Aircraft Accidents. *International Conference on Advanced Computing and Communication Systems*.
 6. Arul Kumar, V & Elavarasan, N. (2014). A Survey on Dimensionality Reduction Technique. *International Journal of Emerging Trends and Technology in Computer Science*, 3(6). ISSN 2278- 6856.
 7. Easha, Mangal & Anurag, Jain. (2016). A Review: Customer Relationship Management Classification using Data Mining Technique. *International Journal of Advanced Research in Computer and Communication Engineering*, 5(2). ISSN 2278- 1021.
 8. Gerard Biau. (2012). Analysis of a Random Forest Model. *Journal of Machine Learning Research*, 13, 1063- 1095.
 9. Gilbert, Nancy & Appavu alias, Balamurugan. (2013). A Comparative Study of Feature Selection Methods for Cancer Classification using Gene Expression Dataset. *Journal of Computer Applications*, 5(3), 0974- 1925.
 10. Girish, Chandrashekar & Ferat, Sahin. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, 16- 28.
 11. Haifeng, Zhao., Zheng Wang. & Feiping, Nie. (2019). A New Formulation of Linear Discriminant Analysis for Robust Dimensionality Reduction, *IEEE Transactions on Knowledge and Data Engineering*, 31(4), 629- 640.
 12. Lakshmi Padmaja, D & Vishnuvardhan, B. (2014). Survey of Dimensionality Reduction and Mining Techniques on Scientific Data. *International Journal of Computer Science and Engineering Technology*, 5(11). ISSN 2229- 3345.
 13. Linton Chandra, Paul., Abdulla Al, Suman., & Nahid, Sultan. (2013). Methodological Analysis of Principal Component Analysis (PCA) Method. *International Journal of Computational Engineering and Management*, 16(2). ISSN 2230- 7893.
 14. Malavika, S & Selvam, K. (2015). Reduction of Dimensionality for High Dimensional Data Using Correlation Measures. *Global Journal of Pure and Applied Mathematics*, 11(1). ISSN 0973-1768.
 15. Mary Walowe, Mwadulo. (2016). A Review on Feature Selection Methods for Classification Task. *International Journal of Computer Applications Technology and Research*, 5(6). ISSN 2319- 8656.
 16. Mehnaz, Khan & Quadri, S.M.K. (2013). Effects using Filter Based Feature Selection on the Performance of Machine Learners Using Different Datasets. *BIJIT-BVICAM's International Journal of Information Technology*, 5(2). ISSN:0973- 5658.
 17. Mohammed, Azam & Mohammed, Tabrez Nafis. (2014). Data Mining of Web Access Logs Using Classification Techniques. *International Journal for Research in Applied Science and Engineering Technology*, 2(8). ISSN 2321- 9653.
 18. Mohana Chelvan, P & Perumal K. (2017). On Feature Algorithms and Feature Selection Stability Measures: A Comparative Analysis. *International Journal Computer Science and Information Technology*, 9(3).
 19. Noelia, Sanchez- Marono., Amparo, Alonso- Betanzos & Maria, Tombilla-Sanroman. (2007). Filter methods for feature selection. A Comparative Study. *International Conference on Intelligent Data Engineering and Automated Learning*, 178- 187
 20. Radhakrishnan, B., Shineraj, G & Anver, Muhammed. (2013). Application of Data Mining in Marketing. *International Journal of Computer Science and Network*, 2(5). ISSN 2277- 5420.

How to cite this article:

Sivakumar Venkataraman *et al.* 2020, Feature Selection Methods: A Survey for effective Data Classification. *Int J Recent Sci Res.* 11(11), pp. 40000-40007. DOI: <http://dx.doi.org/10.24327/ijrsr.2020.1111.5605>
