Original Research Paper

Volume - 13 | Issue - 01 | January - 2023 | PRINT ISSN No. 2249 - 555X | DOI : 10.36106/ijar

Computer Science

AN ANALYSIS OF FEATURE SELECTION TECHNIQUES FOR AN EVOLUTIONARY APPROACH TO GENETIC ALGORITHMS

Dr. V Vasanthi*	Assistant professor, PG & Research, Department of Computer Application Hindusthan College of Arts and Science Coimbatore, India*Corresponding Author
E Vishnu Prasath	Student PG & Research, Department of Computer Application Hindusthan College of Arts and Science Coimbatore, India.
R Manoj Kumar	Student PG & Research, Department of Computer Application Hindusthan College of Arts and Science Coimbatore, India.
R Mukesh	Student PG & Research, Department of Computer Application Hindusthan College of Arts and Science Coimbatore, India.

ABSTRACT Data magnitude is growing expeditiously, which pretends the challenges to extensive majority of current mining and learning algorithms, such as the bane of dimensionality, huge storage requirement, and enormous computational cost. Feature selection has been proven to be an effective and efficient way to prepare high-dimensional data for data mining and machine learning. The recent evolution of novel techniques and new types of data and aspects not only advances existing feature selection research but also emerges the feature selection constantly, becoming applicable to a expansive range of applications. In this entry, we aim to provide a basic introduction to feature selection including basic concepts, classifications of existing systems, recent development, and applications.

KEYWORDS :

1. Introduction

The Feature selection aims at resolving the situation of system misinterpretation by holding the threshold value. The focal point of feature selection is to decide on a rift of data from the input while tumbling effects irrelevant [1].

Feature selection plays a significant role in extracting meaningful information from mountains of data using a minimal subset of attributes.

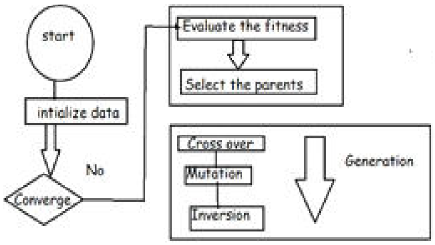
Nowadays, feature selection becomes mandatory as it is difficult to mine and transform the momentous volume of data into valuable insights[1]. The effective use of feature selection can significantly improve the predictive rate of classification of different classifiers.

To deal with very large data instances, bio-inspired approaches such as Genetic algorithm, Particle Swarm Optimization, Cuckoo Search, Artificial Bee Colony algorithms etc have been proposed to overcome the feature selection problem[4].

2. Framework of Feature Selection algorithms

The Objective of the algorithm is minimum amount and maximum quality.

2.1 Framework of Genetic Algorithms



The undeveloped progression adopted by GAs comprises create a preliminary set population and evaluating them. After Selection the predominant solutions are identified with the manipulation of children. The lesser delimiters are identified and eliminated in the process of the selection. The population will be free in dependency anomalies.

3. Feature Selection methodology for Genetic Algorithms

In Developing an application to achieve better results, various feature selection algorithms can be functional and the accurate solution is preferred for the problem. If the algorithm produces a different subset for any perturbations, then becomes it unpredictable for feature selection[6].

3.1 Filter Method

In the Filter method the set of all features are minimized to select a subset, based on which the algorithm is selected and performance is achieved. The initial component is selected based on the available features and the best subset is achieved. Statistical tools such as ANOVA test, Chi-square test and correlation coefficient is done to identify some important features that is correlated with our target[4].

3.2 Wrapper Method

Wrapper method uses the combination of variables to analyse the predictive power. The wrapper method search the amalgamation of variables and it will be computationally expensive then the filter method but it is not compulsory for the high number of features[5].

The types of methods of wrapper method are
1.Subset selection method
2.Forward step selection method
3.Backward step selection method

Forward step Selection method

Forward Selection is an iterative method in which we initialize with null feature added to that the first feature is added to percept whether the desired results is achieved. In each iteration one more input is added to which we get the best solution.

Input	A	B	C	D	E	Check Accuracy
1st iteration	□					Good
2nd iteration	□	□				Good
3rd iteration	□	□	□			Best
4th iteration	□	□	□	□		Not Good

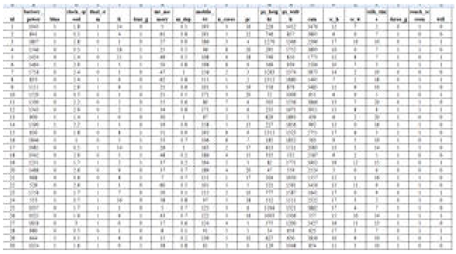
Backward step selection method

Here we have start with the all features and removes the least important t feature after each iteration which improves the model[7]. The table is considered for the example and reverse process is done so that the best model is identified in the second iteration by removing D and E. The Backward Elimination takes the method of chi-square test that take the statistical text and test all the independent variable together if it does not have impact or the low impact.

4. Generating of Quality data with Feature Selection

In the univariant selection, scikit-learn library provides class that can be used with a suite of different statistical tests to select a specific number of features [4].

Consider the Example



The example below uses the chi-squared (chi²) statistical test for non-negative features to select 10 of the best features from the Mobile Price

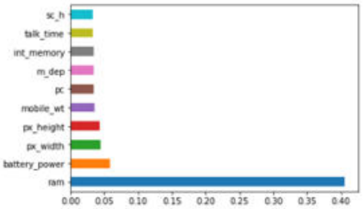
5. Correlation with heat map

Correlation in feature is the process of reducing the number of input variables when developing a predictive model. It is enviable to concentrate the number of input variables to both reduce the computational cost of modeling[3]. In the case if we say independent feature x and dependent feature y, the outputs are correlated with the variables that if x is escalating y is also growing.

Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable[2]. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables.

Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable. Feature importance is an inbuilt class that comes with Tree Based Classifiers, we will be using Extra Tree Classifier for extracting the top 10 features for the dataset[4].

```
Import pandas as pd
import numpy as np
data = pd.read_csv("D://Blogs//train.csv")
X = data.iloc[:,0:20]  #independent columns
y = data.iloc[:,-1]    #target column i.e price range
from sklearn.ensemble import ExtraTreesClassifier
import matplotlib.pyplot as plt
model = ExtraTreesClassifier()
model.fit(X,y)
print(model.feature_importances_) #use inbuilt class feature_importances of tree based classifiers
#plot graph of feature importances for better visualization
feat_importances = pd.Series(model.feature_importances_, index=X.columns)  feat_importances.nlargest(10).plot(kind='barh')
plt.show()
```



6. Genetic Features Selection Algorithms Operators

The probable restriction is that individual feature status may remove potentially useful features without bearing in mind their interactions with other features[4]. Feature ranking approach, where each feature was given a score according to its frequency of appearance in the best GP individuals.

Feature selection was achieved by using only the top-ranked features for classification. This way of evaluating individual features took other features into account, which could avoid the limitation of most single feature ranking methods.

A GP-based multi-objective filter feature selection approach was proposed for binary classification problems[4].

The operators are

1.Encoding

How individual the population representation. It represents the method of representation of genetic algorithms

2.Selection

The selection operator chooses the individual in the population that will create off spring for the next generation

3.Cross over

Combine the parts of parent's chromosomes to create new one

4.Mutation

Randomly invert one gene in chromosome and selecting the features that filter accurate data.

REFERENCES:

1. S. Sreng, N. Maneerat, D. Isarakorn, K. Hamamoto, and R.Panjaphongse, ―Primary screening of diabetic retinopathy based on integrating morphological operation and support vector machine, In 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2017, pp. 250–254. 2.B. Antal and A. Hajdu, ―An ensemble-based system for automaticsc.
2. Antal and A. Hajdu, ―An ensemble-based system for automatic screening of diabetic retinopathy, □ Knowledge-Based Syst., vol. 60, no. April, pp. 20–27, 2014. Agrawal S, Agrawal J (2015) Survey on anomaly detection using data mining techniques.
3. ProcediaComputSci60(1):708–713. https://doi.org/10.1016/j.procs.2015.08.220Ahmed M,
4. Mahmood AN, Islam MR (2016) A survey of anomaly detection techniques in financial domain. https://doi.org/10.1016/j.future.2015.01.001
5. Alelyani S (2013) On feature selection stability: a data perspective. Arizona State University, Tempe Alelyani S, Liu H, Wang L (2011) The effect of the characteristics of the dataset on the selection stability. In: Proceedings—international conference on tools with artificial intelligence,
6. ICTAI, pp 970–977. https://doi.org/10.1109/ICTAI.2011.167,Alelyani S, Tang J, Liu H (2013) Feature selection for clustering: a review. Data Cluster
7. Algorithms Appl29:110–121Alter O, Alter O (2000) Singular value decomposition for genome wide expression data processing and modelling. Proc Natl AcadSci USA 97(18):10101–10106
8. Ambusaidi MA, He X, Nanda P (2015) Unsupervised feature selection method for intrusion detection system. In: Trustcom/BigDataSE/ISPA, 2015 IEEE, vol 1, pp 295–301.