Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2021, 8(7):51-58**

Research Article

ISSN: 2394-2630
CODEN(USA): JSERBR

A Framework for Feature Subset Selection Based on Maximization of Gain Ratio

Olabiyi Olanrewaju Mohammed[1], Bennett, E.O[2], Nwiabu, N.D[3]

Department of Computer Science, Rivers State University, Port Harcourt, Nigeria
E-mails:[1]olabiyilanre@gmail.com, [2]bennett.okoni@ust.edu.ng, [3]nwiabu.nuka@ust.edu.ng

Abstract It is difficult to know which features (often are referred to as meta-feature) are used to characterized a data set. The cost for measuring a feature is a critical issue to be considered when selecting a subset. In the case of emotional speech dataset evaluation, the features may be the number of speakers or accents.Each speaker is associated with its own attribute (emotional states) such as anger, happiness, neutral, sadness, and surprise. The challenge is in selecting the subset of features with minimum cost. So, dimension reduction with the help of Feature Subset Selection (FS) is a useful tool. This paper presents a framework for feature subset selection based on maximization of gain ratio. The gain ratio has been applied to resolve the problem of choosing redundant and irrelevant features in certain circumstances. However, the reduced dataset was further processed using Bayesian classifier to minimize the probability of classification error under the assumption that the sequence of points is independent. Gain ratio with feature selection resulted in features subset of which are given as input to Bayesian classifier. Result shows that the accuracy obtained after classification was 82.50% which is significantly appropriate.

Keywords Feature Selection, Feature Subset Selection, Gain Ratio, Bayesian Classifier

1. Introduction

In the fields of data mining and machine learning, Feature Subset Selection (FSS) is crucial. A good FSS algorithm can efficiently eliminate irrelevant and redundant features while also taking feature interaction into consideration. This not only increases a learner's performance by increasing the generalization capacity and interpretability of the learning model, but it also leads to a better understanding of the data [1]. Many areas related to expert and intelligent systems use feature selection, including data mining and machine learning, image processing, anomaly detection, bioinformatics, and natural language processing. Due to its computational efficiency, scalability in terms of dataset dimensionality, and independence from the classifier, feature selection based on information theory is a popular approach.

Feature extraction and feature selection are the two major types of dimensionality reduction techniques. Existing features are transformed into a new feature space with a lower dimensionality using feature extraction techniques. Because feature selection does not change the data, it is the best option when a thorough understanding of the underlying physical process is required. When only discrimination is required, feature extraction may be preferable [2]. In embedded processes, the feature selection and learning stages are combined. These methods are less computationally costly and less prone to overfitting; however, they are limited in their generalization because they are very specific to the learning algorithm used [3].

There has been a recent increase in the number of features collected and stored in databases, but many of these features are irrelevant or redundant. These features not only have no place in the knowledge discovery process, but they also add to the findings' complexity and incomprehensibility. However, determining which features

(often referred to as meta-features) are used to characterize a data set can be difficult. The cost of measuring a feature is an important factor to consider when choosing a subset. The number of speakers (1,2,3,4) or accents ((Boston, General, and New York)) may be used to evaluate emotional speech datasets. Each speaker has its own emotional state (attribute), such as anger, happiness, neutral, sorrow, and surprise. The difficulty lies in determining which features to include in the subset with the least amount of risk. As a result, dimension reduction using Feature Subset Selection (FS) is a beneficial step.

The Gain Ratio (GR) is a method of reducing bias in information gain. By using intrinsic information from each attribute, the gain ratio increases information gain. Because E(S) is constant for all characteristics A, increasing information gain is equivalent to reducing average entropy [4]. When choosing an attribute, the gain ratio considers the number and size of branches. By taking into account the inherent information of a split, it corrects the information gain. The entropy of instance distribution into branches is intrinsic information; how much information do we need to know which branch an instance belongs to, as the amount of intrinsic information increases, the value of the attribute decreases. Gao et al. in [5] considered feature selection to be a software engineering search problem.

This main focus of the paper is to develop a framework for feature subset selection based on maximization of gain ratio. This method will reduce the dimensionality of the data and keep the number of features as low as possible, in order to decrease the training time and enhance the classification accuracy of the algorithm.

2. Literature Review

Feature selection is a method for selecting the best subset of features based on a set of criteria. Importance of feature selection: to enhance the model's performance (when it comes to speed, predictive power, and model simplicity), to minimize dimensionality and noise, and to visualize data for model selection

In software engineering, feature selection is regarded as a search problem. They proposed a hybrid Firefly Search (FS) method based on the Kolmogorov–Smirnov statistic and Automatic Hybrid Search for software quality estimation in their research (AHS). Their findings revealed that removing 85 percent of software metrics had an effect on results [5].

In [6] conducted a large-scale impact analysis on twenty-one widely used classifiers using twenty-eight Firefly Search (FS) techniques. NASA's software deficiency datasets and the PROMISE repositories were used in their experiment. They found that the correlation-based filter-FS approach, which is based on the Best First (BF) search method, outperforms other FS methods across datasets. This indicates that they used a wide range of FS approaches, classification methods, and datasets in their study. They discovered the Best First (BF) and Genetic Algorithm search methods as search mechanisms for the FSS methods. Heuristic and meta-heuristic search methods include Bat Search (BAT), Ant Search (AS), and FS.

A standard study by empirically comparing cutting-edge Firefly Search (FS) methods in [7]. All of the following methods have been considered: IG, RF, PCA, Correlation-based Feature Subset Selection (CFS), Consistency Feature Subset Selection (CNS), and Wrapper Subset Evaluation (WRP). Five software defect datasets were used to test NB and DT, and the predictive models were tested using the Area Under Curve method (AUC).Their findings revealed that while FS is beneficial to SDP, there is no single best FS method for Software Deficiency Prediction (SDP). This may be due to the number and types of software defect datasets considered, as well as the positive or constant search procedures used in the FSS and WRP FS techniques of Software Deficiency Prediction (SDP) models.

On the basis of FFS on SDP, in [8] conducted a comparative analysis of classifiers. Their findings credited the use of FFS, but other FS approaches may still be used for further study. Wrappers produce better-performing subsets than filter feature selection, as shown by the detail that the subsets were tested using a real-world modeling procedure.

On four software defect datasets, in [9] based on three distinct FFR and WRP models, conducted comparative studies on Firefly Search (FS) techniques. Their findings revealed that smaller data sets can retain predictability by having less features than larger data sets.

Kohavi and John in [10] defined domains in which a function appears in the target concept to be learned but not in the optimal feature subset that maximizes the predictive accuracy for the learning algorithm in use. This is

due to the classifier's inherent characteristics and limitations: function importance and accuracy optimality are not always correlated in FSS.

When the aim of FSS is to achieve the highest level of accuracy, the features selected should be based on the learning algorithm as well as the features and target principle to be learned, according to John et al. in [11].

The embedded feature selection technique introduced in [12]. Feature selection is achieved indirectly during the construction of the classification algorithm in this definition.

A classification system for feature selection algorithms was given in [13]. First, approaches focused on computational pattern recognition methods and those based on artificial neural networks are separated. Methods that guarantee finding the optimal solution are separated from algorithms that which result in suboptimal feature sets in the statistical pattern recognition group. The suboptimal methods are further categorized into processing and manipulating a single feature subset and methods that operate with a populace of subsets. There is a modification made between deterministic and nondeterministic algorithms for each of these and stochastic methods. For each execution of a deterministic algorithm, the same result is produced for a given problem. Since stochastic methods use a random variable, they produce different subsets on separate runs.

3. System Design

Structural details of the system including the component of the feature selection framework. Figure 1 shows the feature selection framework, whose components are heuristic search and gain ratio.

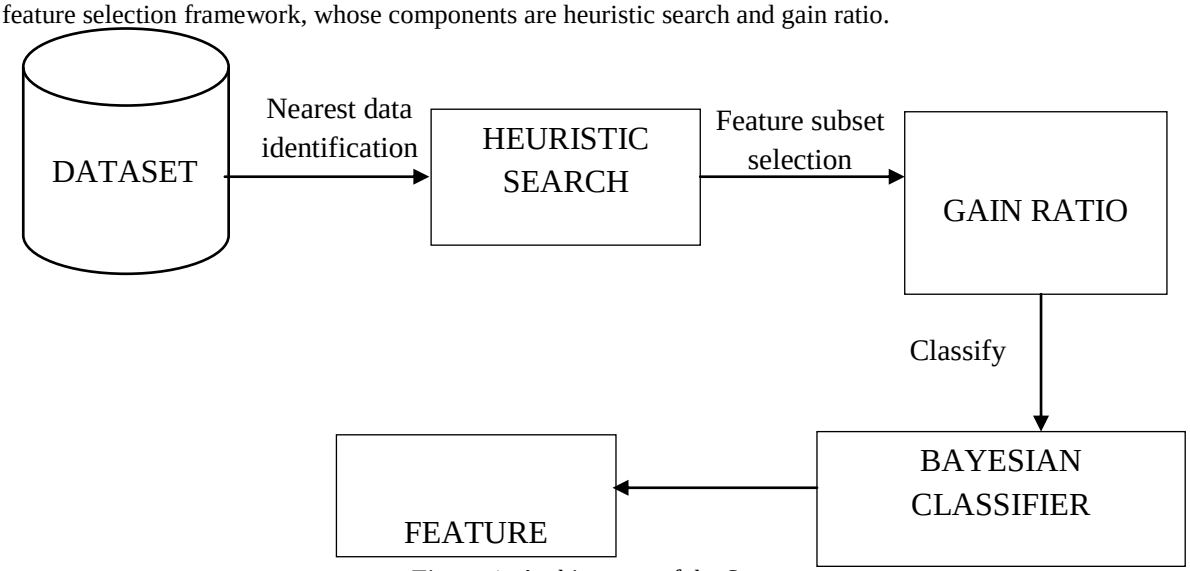

Figure 1: Architecture of the System

Gain ratio takes number and size of branches into account when selecting an element. Bayesian classifier has been chosen due to the nature of the instance-based learning technique and its approach. The idea behind a Bayesian classifier is that, if the class is identified, the values of another feature can be projected. If the class is not identified, Bayes' rule can be utilized to envisage the class given the attribute values. In a Bayesian classifier, the learning agent constructs a probabilistic model of the features and uses that model to envisage the cataloging of a new instance.

3.1. Dataset

As the basis of this findings, we have used the emotional speech dataset that has been studied as a machine learning dataset. The emotional speech which is Danish Emotional Speech (DES) consists of number of classes (N) = 5042 utterances expressed by 4 actors under number of samples (C) = 5 emotional states, such as surprise, happiness, sadness, neutral and anger. Data from 9 speakers with 3 regional accents (Boston, General, and New York) are exploited.

Dataset entails of the large features of a set of datasets. Let D = {a1, a2, a3, a4} be a sequence of N datasets. I, F, T and D. Thus, dataset dimensionality, DD = I/F.

I= the amount of instances in *D.*

F=the amount of features in *D*.
T= the amount of concept values in *D*.
Thus, D is the dataset.

3.2. Nearest Data Identification using Heuristic Search

Heuristic Search is a form of search that makes use of heuristics. Heuristic Search, like a depth-first search, shows a path linking the beginning and end graph. This path's maximum length is M, and the amount of subclasses generated is O. (M).In order to discover a nearer ideal subset of features in a quicker process, the heuristic must be chosen carefully.

Considering a subset graph formed with nodules matching to column subclasses. There is an edge from subclass S_i to subset S_j if including one column to *Si* generates *Sj*. The graph generated for the matrix A = (a1, a2, a3) as shown in Figure 2.

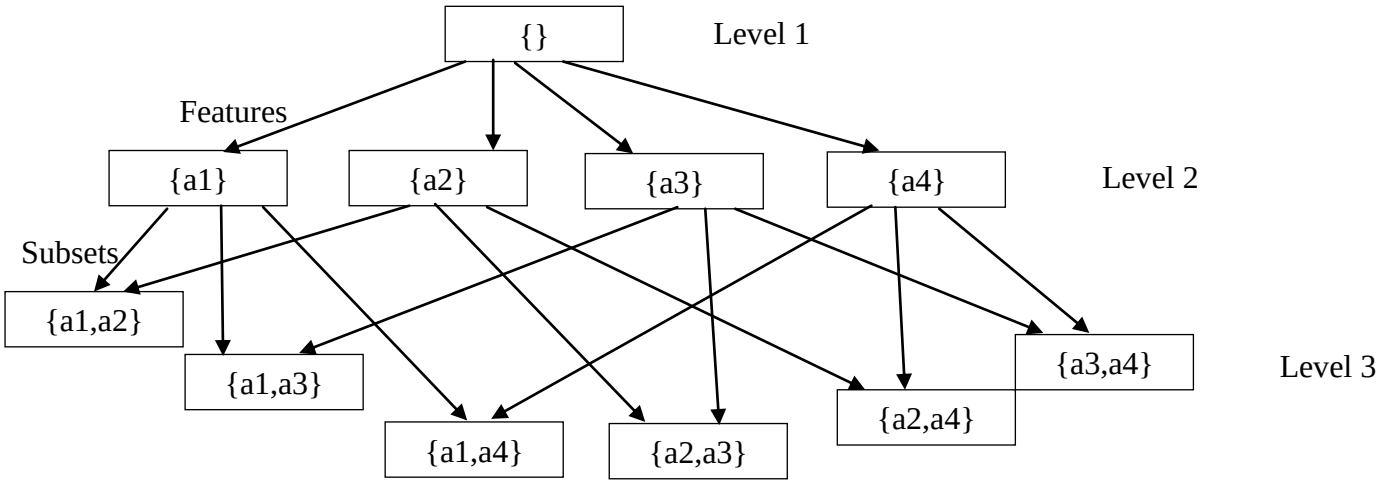

Figure 2: Feature Subset

Despite the fact that a subset graph is not a tree, it is made up of two assets that are common to trees. The center, which corresponds to the empty subset, is the first property. The second point to note is that all routes from the root to a nodule are equal. For instance, if the accurate nodule {a2, a4} is established, it is irrelevant if it is reached by the route {} -> {a2} -> {a2, a4} or by the route {} -> {a4} -> {a2, a4}. This is alike to the situation of a tree where the optimal of route leading to a nodule is irrelevant as there is a distinctive route leading from the root to any nodule.

The k nearest datasets of {a1}are known by computing the distance among {a1}and each dataset based on their attributes. The lesser the gap, the more alike the resultant data to {a1}.

Let $f_i = \{f_{i,1}, f_{i,2}, \dots, f_{i,h}\}$ be the features of dataset Di, where $F_{i,p}$ is the value of P_{th} attribute of F_i and *h* is the length of the attributes. The L_1 norm distance among data a_i and a_j can be articulated as:

$$dist(a_i, a_j) = f_i - f_j = \sum_{p=1}^{h} f_{i,p} - f_{j,p} \quad (1)$$

Where,
a_i= data *i*
a_j= data *j*
f_i= feature *i*
f_j= feature *j*
h = length of feature.
However, *i* and *j* are different number in a given dataset.
length is the sample size of dataset.

Procedures for Heuristic Search:

1. Put the root node into F

2. while F is non-empty.
3. pick a_i with the smallest f'(a_i) from F.
4. If a_i has k columns return it as the solution
5. Else
6. Adda_i to C
7. Examine all children a_j of a_i
8. If a_j is in C or F do nothing.
9. Else
10. Put a_j in F

3.3. Gain Ratio for Feature Subset Selection and Ranking

In this feature selection approach, the info Gain Ratio (GR) is formulated for each attribute of the training dataset *D* to find the major attribute based on the info existing in the features of the *D* [14].

Let *S* be sequence comprising of *D* data models with *m* different classes. The anticipated info needed to categorize a given sample is given by:

$$Entropy(S) = \sum_{i=1}^{m} p_i \, log_2(p_i) \quad (2)$$

Entropy offers the info needed in bits; this can consist of fractions of bits. Thus, info is measured in bits.

Where,

p_i= the likelihood that a random sample fits to class C_i.

m = Number of partitions S

S = Set of cases

Cases are instances, that is attributes with a large number of values.

Let element *A* has *v* different values. Let S_{ij} be amount of samples of class Ci in a subclass *Sj*.*Sj* encloses those models in *D* that have value *aj* of *A*. The entropy, or anticipated info based on the segregating into subclasses by *A*, is given by:

$$E(A) = -\sum_{i=1}^{m} I(D) \frac{|s_1 + s_2 + \cdots s_m|}{|S|} \quad (3)$$

D = Overall dataset (dataset is a collection of data).

A = Subset attribute (attribute is a property of an element).

m = Amount of feature partition A

| Si | = Subset dimensions of dataset owned feature on A partition

| S | = Total amount of cases in dataset

However,

Subset attribute is the subgroup of elements defined by attributes of the same kind.

Amount of feature partition Ais the number of total possible partitions of an A set.

Total amount of cases in a set of data is the sum amount of attributes of the same kind in a dataset.

The info that would be gained by branching on *A* is

Gain (A) = I(D) – E(A)

gain ratio which applies normalization to information gain utilizing a value defined as

$$SplitInfo_A(D) = -\sum_{i=1}^{v} \left(\frac{|D_i|}{|D|}\right) log_2 \left(\frac{D_i}{D}\right) \quad (4)$$

The info produced by dividing the training data set *D* into *v* partitions conforming to *v* products of a test on the feature *A*.

Where,

D = Total set of data

A = Subset feature

v = Amount of partition features A

| Di | = Subset dimensions of dataset owned feature on A partition

| D | = Total amount of cases in dataset

The gain ratio is formulated as:

Gain Ratio (A) = Gain (A)/ SplitInfo (A)

4. Results and Discussion

Danish Emotional Speech (DES) corpus is deployed for feature subset selection. Different emotional states are found in the corpus, such a ssurprise, happiness, sadness, neutral and anger. These emotional states are five, thus, all utterances corresponding to these five emotional states are used for selection of feature, since the target is to achieve the maximum recognition rate for emotional state classification with the selected feature subsets. Table 1 to Table 1.2 shows the 90 features generated from the utterances that embrace the mean, median and variance of pitch, energy contours and formants. Instead of directly handling data with the collection of features to the learning procedure after generating features, selection of feature for cataloguing conducted feature selection to pick a subset of features and then processed the data with the chosen features to the learning algorithm.

In Table 1, and according to Table 1.1 and Table 1.2 much useful information is obtained. The classification reached its peak value when 8 features subset are selected. The feature subsets are (1,2,3,4,5,16,21,46), using the gain ratio to rank, the system selected 8 features subset out of 90 generated features.Thecorrect classification rate ranges from 0.205 to 0.310. The observation that the correct classification rate and the number of correctly categorized utterances are closely related led to an estimation of the variance of the correct classification rate. Obviously, the classifier's preference has little effect on the variance of the proper classification rate. The execution time was 765.766 seconds.

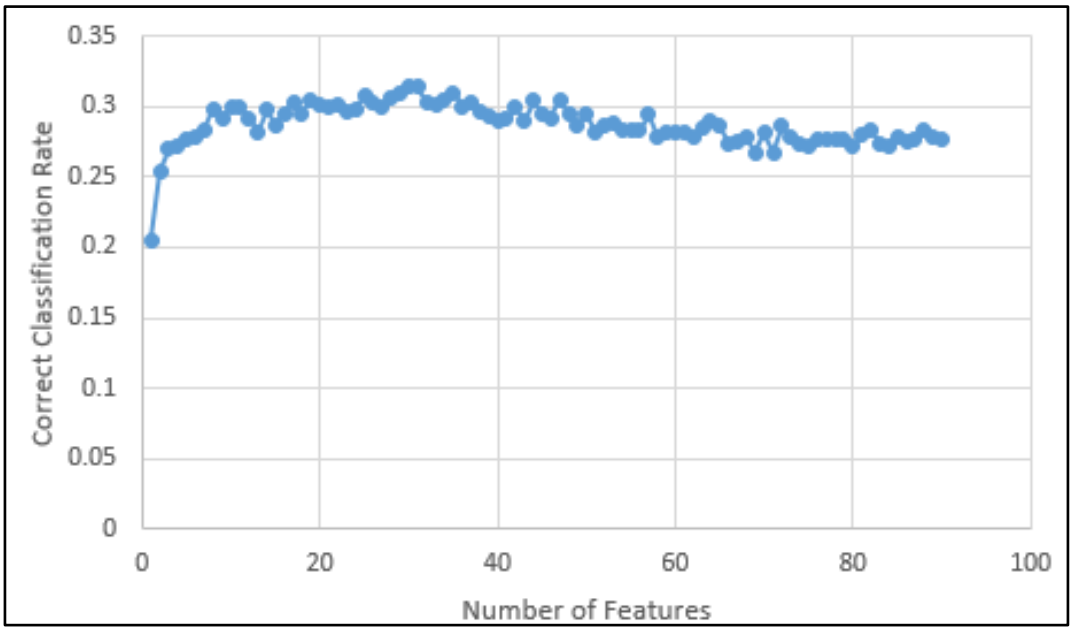

Figure 3: Correct classification rate against correct classification

Figure 3 shows the cross classification rate against correct classification. The accuracy was 82.50 percent with 5 attributes (four input and one output type attribute). The pruned identified indignation, pleasure, neutrality, depression, and surprise as significant qualities a number of times. The regular K cross validation method was utilized, with K = 5. For the benefit ratio, the size is 90, the sum of generations is 90, the right classification is between 0.275 and 0.300, the device conducted 5 repetitions during a preliminary function test, and the best current is 0.289. The mix of gain ratio and function collection resulted in a subset of features that were fed into the Bayesian classifier. K fold cross validation with K = 5 and complete data as training data were used for the gain ratio process. The function subsets were produced in 765.766 seconds, indicating that the system is sluggish.

The Danish emotional speech dataset is of high dimension. Gain ratio has been used as ranking method to choose a subset of 5 attribute from the dataset.In reduced subset of Danish emotional speech dataset considering

anger, happiness, neutral, sadness, and surprise. Table.1 shows the accuracy achieved after classification using the Bayesian Classifier algorithm.

Table 1: Classification accuracy (%) on Danish dataset for five classes (k=5): angry, happy, sad, neutral and surprises

Gain Ratio Reduced Dataset	Classification Accuracy %
K=3	73.83
K=5	82.50

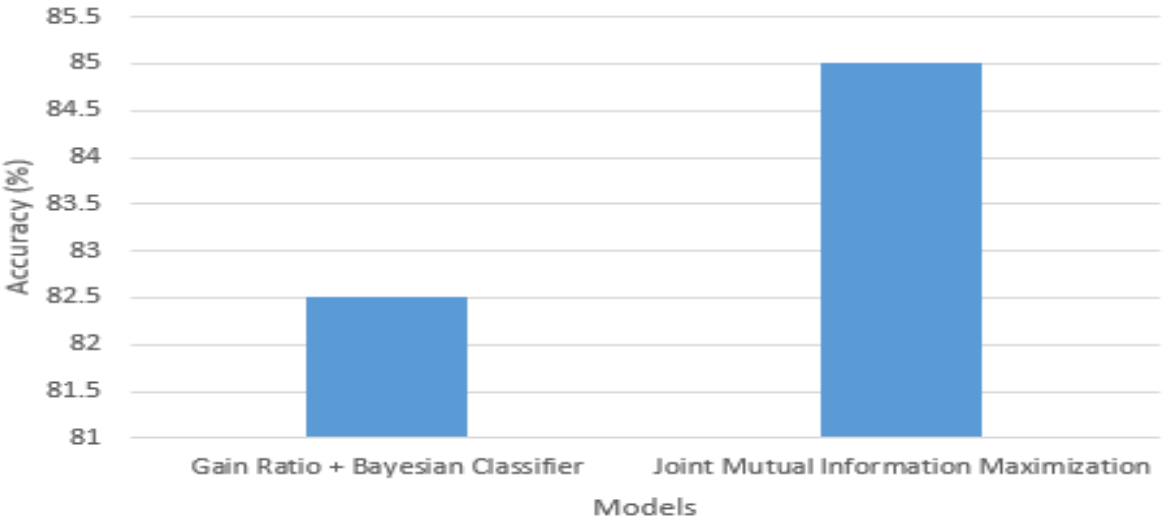

Figure 4: Comparison of Gain Ration Reduced Dataset and Joint Mutual Information Maximization

The accuracy for gain ratio reduced dataset for Bayesian classifier on Danish emotional dataset is 82.50%. For the joint mutual information maximization, the accuracy is 85% as illustrated in Table2.

Table 2: Comparison of Gain Ration Reduced Dataset and Joint Mutual Information Maximization

Models	Accuracy %
Gain Ratio + Bayesian Classifier	82.50
Joint Mutual Information Maximization	85

5. Conclusion

The aim of selection of feature is to prevent choosing too many or too few features. If very few features are chosen, the information quality in this set of features is likely to be limited. However, if too many (irrelevant) features are chosen, the effects of noise in (most real-world) data can overshadow the info present. As a consequence, every function selection process must resolve this tradeoff. On the basis of the Danish emotional expression corpus, we focused on a function subset selection strategy. Gain ratio was used in this analysis to rate the attributes of the datasets used. The gain ratio approach is meant to address the case of picking redundant and irrelevant features under some situations. However, under the assumption that the sequence of points is distinct, the reduced dataset was further analyzed using a Bayesian classifier to reduce the likelihood of classification error. A Bayesian classifier works on the principle: if a class is defined, the values of other features can be estimated. Bayes' rule can be utilized to estimate the class provided (some of) the function values if the class is unknown. The learning agent in a Bayesian classifier creates a probabilistic model of the features and uses it to forecast the cataloguing of a new case. The system identifies language independent feature subsets for emotional speech, top 8 feature subsets were selected and Reduction of high dimensionality on dataset.

References

[1]. Zhao, Z. and Liu, H. (2007) Semi-Supervised Feature Selection via Spectral Analysis. In Proceedings of SIAM International Conference on Data Mining, Philadelphia, PA: Society for Industrial and Applied Mathematics, 641-646.

[2]. Jain, A. K., Duin, R. P. W. and Mao, J. (2000). Statistical Pattern Recognition: A Review. IEEE Transaction Pattern Analysis and Machine Intelligence, 22, 4–37.

[3]. Guyon, I., Gunn, S., Nikravesh, M. and Zadeh, L. A. (2006). Feature extraction foundations and applications. Springer Studies in fuzziness and soft computing.

[4]. Han, J. and Kamber, M. (2001). Data Mining Concepts and Techniques, Morgan Kaufmann.

[5]. Gao, K., Khoshgoftaar, T. M., Wang, H. and Seliya, N. (2011). Choosing software metrics for defect prediction: an investigation on feature selection techniques. Software Practical Experiment, 41, 579–606.

[6]. Ghotra, B., McIntosh, S. and Hassan, A. E. (2017). A Large-Scale Study of the Impact of Feature Selection Techniques on Defect Classification Models. In Proceedings of the 2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR), 146–157.

[7]. Afzal, W. and Torkar, R. (2016). Towards Benchmarking Feature Subset Selection Methods for Software Fault Prediction. In Computational Intelligence and Quantitative Software Engineering, 33–58.

[8]. Akintola, A. G., Balogun, A. O., Lafenwa-Balogun, F. B. and Mojeed, H. A. (2018). Comparative Analysis of Selected Heterogeneous Classifiers for Software Defects Prediction Using Filter-Based Feature Selection Methods. FUOYE Journal of Engineering and Technology, 3, 134–137.

[9]. Rodriguez, D., Ruiz, R., Cuadrado-Gallego, J., Aguilar-Ruiz, J. and Garre, M. (2007). Attribute Selection in Software Engineering Datasets for Detecting Fault Modules. In Proceedings of the 33rd EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO). IEEE, 418–423.

[10]. Kohavi, R. and John, G. H. (1997). Wrappers for Feature Subset Selection. Artificial Intelligence, 97(1-2), 273-324.

[11]. John, G., Kohavi, R. and Pfleger, K. (1994). Irrelevant features and the subset selection problem, in: Proceedings 11th International Conference on Machine Learning, 121–129.

[12]. Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. Artificial Intelligence, 245–271.

[13]. Zongker, D. and Jain, A. (1996). Algorithms for feature selection. International Conference on Pattern Recognition, (ICPR), 18–22.

[14]. Uguz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-based Systems, 24(7), 1024–1032.