Journal of Computer Science 7 (5): 800-820, 2011
ISSN 1549-3636

Feature Selection for High Dimensional Data: An Evolutionary Filter Approach

[1]Anwar Ali Yahya, [1]Addin Osman, [2]Abd Rahman Ramli
and [1]Adlan Balola
[1]Faculty of Computer Science and Information Systems,
Najran University, Kingdom of Saudi Arabia
[2]Intelligent Systems and Robotics Laboratory
Institute of Advanced Technology, University Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia

Abstract: Problem statement: Feature selection is a task of crucial importance for the application of machine learning in various domains. In addition, the recent increase of data dimensionality poses a severe challenge to many existing feature selection approaches with respect to efficiency and effectiveness. As an example, genetic algorithm is an effective search algorithm that lends itself directly to feature selection; however this direct application is hindered by the recent increase of data dimensionality. Therefore adapting genetic algorithm to cope with the high dimensionality of the data becomes increasingly appealing. **Approach:** In this study, we proposed an adapted version of genetic algorithm that can be applied for feature selection in high dimensional data. The proposed approach is based essentially on a variable length representation scheme and a set of modified and proposed genetic operators. To assess the effectiveness of the proposed approach, we applied it for cues phrase selection and compared its performance with a number of ranking approaches which are always applied for this task. **Results and Conclusion:** The results provide experimental evidences on the effectiveness of the proposed approach for feature selection in high dimensional data.

Key words: Genetic algorithm, feature selection, high dimensional data, filter approach, Machine Learning (ML), evaluation function, proposed approach, search algorithm, natural language processing, mutation operator

INTRODUCTION

Machine Learning (ML) is a rapidly expanding field with many applications in diverse areas such as natural language processing (Marquez, 2000), bioinformatics (Baldi and Brunak, 2001), image processing (Sajn and Kukar, 2010; Lee *et al*., 2010). It provides tools by which large quantities of data can be automatically analyzed. Fundamental to ML is feature selection, also called dimensionality reduction, which identifies the most salient features, so that the ML algorithm focuses on data aspects most useful for analysis and future prediction. Feature selection algorithm repeatedly selects a subset of original features called candidate subset and measure the optimality of the candidate subset using evaluation function. In doing so, the feature selection approach reduces data dimensionality, removes irrelevant data, increases learning accuracy and improves result comprehensibility (Blum and Langley, 1997; Dash and Liu, 1997; Kohavi and John, 1997).

Technically speaking, feature selection algorithm consists of four basic processes, shown in Fig. 1: subset generation, subset evaluation, stopping criterion and result validation (Liu and Yu, 2005). These steps are performed by three core components, namely, search algorithms, evaluation function and performance analyzer (Dash and Liu, 1997). Subset generation is performed by the search technique (Blum and Langley, 1997) that repeatedly generates candidate feature subset and evaluates it using the evaluation function until a given stopping criterion is met. The selected best subset usually needs to be validated by the performance analyzer, usually by applying the ML algorithm on new instances of data using the selected features. Feature selection is considered successful if the dimensionality of the data is reduced and the performance of the ML improves or remains unaffected.

Corresponding Author: Anwar Ali Yahya, Faculty of Computer Science and Information Systems, Najran University, Kingdom of Saudi Arabia

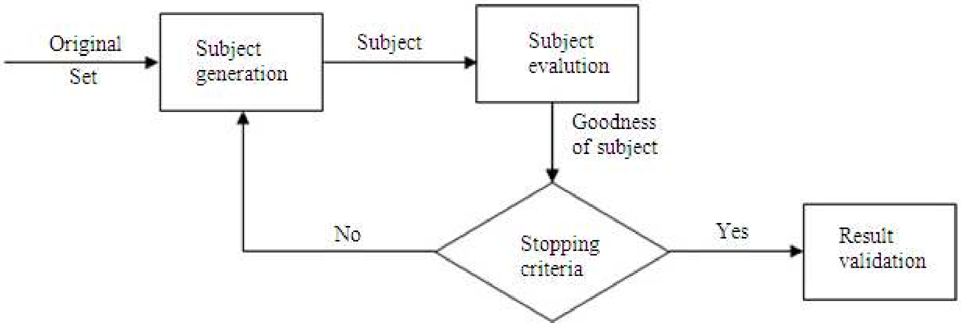


Fig. 1: Feature selection process (Liu and Yu, 2005)

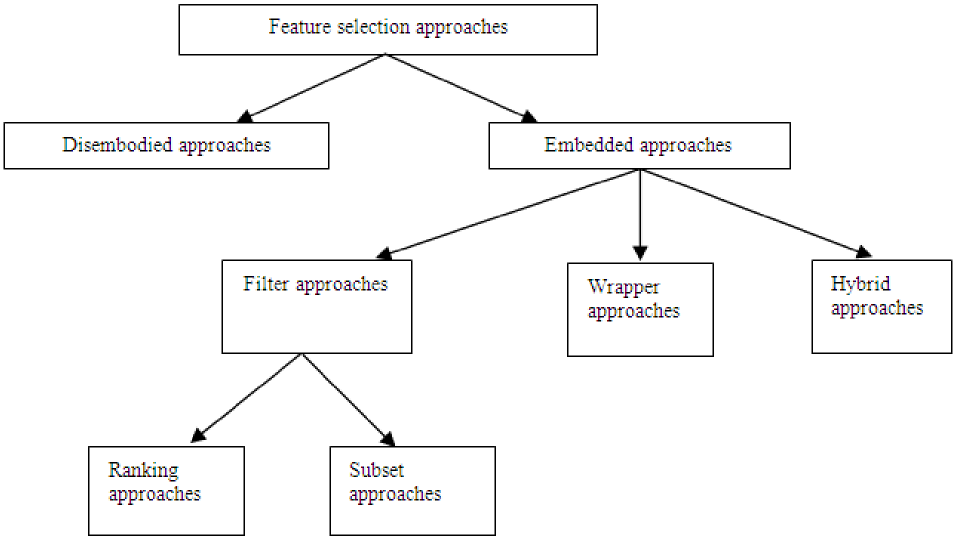


Fig. 2: Taxonomy of feature selection approaches

Feature selection has been a fertile field of research and development since the 1970s in statistical pattern recognition (Ben-Bassat, 1982; Siedlecki and Sklansky, 1988) and ML (Blum and Langley, 1997; Kohavi and John, 1997; John *et al*., 1994). As a result, various feature selection approaches have accumulated over the years. To better underst and the inner instrument of each approach and the commonalities and differences between them, several taxonomies have emerged such as those proposed by Dash and Liu (1997), (Liu and Yu, 2005; Saeys *et al*., 2007). This study adopts the taxonomy depicted in Fig. 2, which has gained a wide consensus among researchers.

At the top level, the taxonomy splits the feature selection approaches into two categories, namely embedded and disembodied, based on whether the feature selection process is incorporated into the ML process of model construction or performed separately. The embedded feature selection approaches (Lal *et al*., 2006) search for the optimal subset of features during the model construction and can be viewed as a search in the combined space of features and models. Embedded approaches are thus specific to a given ML algorithm and consequently have the advantage of including the interaction with the constructed model, while at the same time being computationally feasible. Conversely, the disembodied approaches perform the feature selection as a separate process before the application of ML algorithm.

At the second level, the disembodied feature selection approaches are classified as filter, wrapper, or hybrid, based on the type of the evaluation function. Wrapper approaches use the ML algorithm itself to evaluate the goodness of feature subset. The rationale is that the ML algorithm that will ultimately use the feature subset should provide a better estimation of the goodness of the feature subset (Langley, 1994). Among the advantages of wrapper approaches are their selection is aware of the inductive bias of ML and the ability to take into account feature dependencies between the selected features. Common drawbacks of these approaches are the higher risk of over fitting and the expensive cost of the required computation. Filter approaches, on the other h and, assess the goodness of feature subset by using independent measure, which is based on the intrinsic properties of the data, rather than the ML algorithm. In subsequent stage, the selected subset of features is presented as input to the ML algorithm. In this regards, a wide variety of measures have been used as evaluation function for filter approaches. To cite but few, consistency driven measure (Almuallim and Dietterich, 1991), information theoretic measures (Ben-Bassat, 1982), dependency measures (Hall, 2000), consistency measures (Liu and Motoda, 1998) and even using another ML algorithm as a filter (Hall, 1999). Some of the advantages of filter approaches can be summarized as that they are easily scaled to very high dimensional data and they are computationally simple and fast. A common drawback of filter methods is that they ignore the interaction with the constructed model. Hybrid approaches were devised to alleviate the time complexity imposed by the use of wrapper approaches by hybridizing them with filter approaches. The hybrid approaches combine filter and wrapper approaches to achieve best possible performance with a particular ML algorithm (Xing *et al*., 2001; Das, 2001).

As depicted in Fig. 1, the filter approaches themselves are further partitioned into two groups, namely, ranking approaches and subset search approaches, based on whether they evaluate the goodness of features individually or through feature subsets. Ranking approaches assign weights to features individually based on their in formativeness to the target concepts. A well known example of ranking approaches is Relief (Kira and Rendell, 1992). The main drawback of Kira and Rendell approaches is that they can only capture the relevance of features to the target concepts, but cannot discover dependencies between them. Conversely, subset search approaches, such as FOCUS (Almuallim and Dietterich, 1991), employ search algorithm to search through candidate feature subsets (Dash and Liu, 2003). The search is guided by certain evaluation to capture the goodness of each subset and ultimately an optimal (or near optimal) subset is selected when the search stops (Liu and Motoda, 1998). Unlike the ranking approaches, feature subset approaches evaluate feature as a whole and therefore take into account the dependency between selected features.

In spite of the vast body of feature selection approaches, the incessant increase of data dimensions (number of features) and data size (number of instances) poses sever challenges with respect to their efficiency and effectiveness (Liu and Yu, 2005; Zheng and Zhang, 2007). One of these challenges, which is that the focus of this study, is the so-called curse of dimensionality (Hastie *et al*., 2001). Classically, the dimensionality is considered low if the number of features is of some tens and high if the number is in the range 100-500 (Moser and Murty, 2000). However, in recent applications such as natural language processing, genome analysis and astronomy, the dimensionality of the data can be thousands and even tens of thousands. Such high dimensional data causes a major problem for the feature selection approaches as most of these approaches have quadratic or higher time complexity about the data dimensionality, which consequently affect their efficiency. In view of the above taxonomy, it is not hard to conclude that the wrapper approaches are impractical for such data, due to the time complexity of using the ML algorithm as the evaluation function and the complexity of searching the huge search space. The embedded approaches, on the other h ands, are specific to some ML algorithms, though, their time complexity is far less than wrapper approaches. It is, therefore, commonly accepted fact that the filter approaches are preferred for feature selection in high dimensional data due to their computational efficiency (Liu and Yu, 2005; Zheng and Zhang, 2007; Duch, 2006). Some examples of researches that use filter approach for feature selection in high dimensional domains are (Biesiada and Duch, 2005; Bins and Draper, 2001; Yu and Liu, 2003; Li *et al*., 2004; Guo *et al*., 2008). Within the filter approaches, the subset search approaches are more efficient than the ranking approaches, due to the inability of the ranking approaches to account for the dependencies between the selected features.

Although subset search approaches sound the most suitable, among others, for feature selection in high dimensional data, the scalability of these approaches is affected drastically as the dimensionality of data becomes high. In order for these approaches to cope with the high dimensions of the data, either some simplification assumptions are adopted, or an adapted version of these techniques has to be developed. Genetic Algorithm (GA) is a striking example of subset search approaches that have been applied successfully for feature selection in various contexts (Liu *et al*., 1995; Ozdemir *et al*., 2001; Zhang and Hu, 2005; Lanzi, 1997), due to its advantages over many other search approaches when search spaces are highly modal, discontinuous, or highly constrained (Zhu *et al*., 2006). Despite the striking success of GA, the increase of the data dimensionality poses a challenge to its straightforward application and consequently to its efficiency (Hong and Cho, 2006). Some attempts to address this challenge have been made by introducing a simplification assumption with respect to the number of features that must be selected (Liu and Yu, 2005; Hong and Cho, 2006; Sanchez-Ferrero and Arribas, 2007; Lecocke and Hess, 2007). Definitely, such assumption is not correct as the number of features that must be selected cannot be known a priori.

In this study, instead of assuming that the number of the selected features is known a priori, an adapted version of genetic algorithm for feature selection in high dimensional data is developed. The adapted version exploits the variable length representation scheme, hence called Variable Length Genetic Algorithm (VLGA) and makes use of a set of genetic operators to genetically manipulate the variable length chromosomes.

Genetic algorithm for feature selection: GA is a biologically inspired search algorithm, which is loosely based on molecular genetics and natural selection. The basic principles of GA were stated by Hong and Cho (2006). Since then, GA has been reviewed in a number of works (Goldberg, 1989; Haupt and Haupt, 2004; Mitchell, 1996; Vose, 1999). In the standard GA, the candidate solutions are described as bit strings (referred to as chromosomes) whose interpretation depends on the application. The search for an optimal solution begins with a r random population of initial solutions. The chromosomes of the current population are evaluated relative to a given measure of fitness, with the fit chromosome selected probabilistically as seeds for the next population by means of genetic operations such as random mutation and crossover. In general the standard GA consists of the following components.

Population: It consists of a predefined number of chromosomes, in which each chromosome represents a potential solution of a given problem.

Fitness function: It is the driving force of the evolution in GA. The fitness function returns a numerical value, which is supposed to be proportional to the utility or the ability of the potential solution that chromosome represents.

Selection scheme: It is used to select a chromosome in the population for genetic operations. It is based on the survival-of-the-fittest strategy.

Genetic operators: They are the basis of the GA evolution. They recombine the chromosomes of the current population to produce a new population. Conventionally, three operators are implemented; reproduction, crossover and mutation. In the reproduction operator, a chromosome is r randomly selected from the current generation based on its fitness and then copied without any change into the next generation. The crossover probabilistically selects two chromosomes from the current population based on their fitness values and then recombines them to generate offspring. The mutation operator insures the population against permanent fixation by randomly flipping the bits value of a selected chromosome at a randomly selected position.

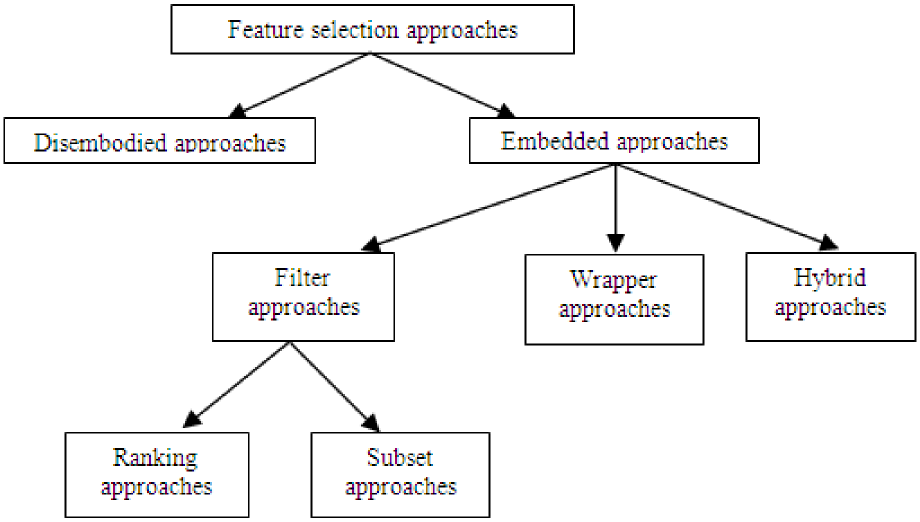


Fig. 3: Standard GA procedure

Stopping criteria to decide when to terminate the run of GA and the control parameters are the probabilities values that control the execution GA. The general procedure of the standard GA is given in Fig. 3.

The seminal work on using GA for feature selection goes back to Siedlecki and Sklansky (1988). Since then, there have been numerous works on using GA for feature selection in various contexts, in wrapper (Liu and Yu, 2005), filter, or hybrid mode. As previously mentioned, in the wrapper mode the ML algorithm is used as the evaluation function, therefore a brief review of the works of GA in the wrapper mode can be carried out based on the employed ML algorithm. K-nearest neighbor was the first ML algorithm employed as GA fitness function in the seminal work of Siedlecki and Sklansky (1988). Kelly and Davis (1991) used GA with k-nearest neighbor to find a vector of weightings of the features to reduce the effects of irrelevant or misleading features. Similarly, GA was combined with k-nearest neighbor to find an optimal feature weighting to optimize a classification task (Punch *et al*., 1993). This approach has proven especially useful with large data sets, where standard feature selection techniques are computationally expensive. GA with fitness function based on the classification accuracy of k-nearest neighbor and features subset complexity was used to improve the performance of image annotation system (Lu *et al*., 2008). Li *et al*. (2001) combined GA and k-nearest-neighbor to select feature (genes) that can jointly discriminate between different classes of samples (e.g. normal versus tumor). This approach is capable of selecting a subset of predictive genes from a large noisy data for sample classification.

Artificial neural networks were employed as GA fitness function for feature selection in several works. For example, GA with neural networks was combined for feature selection in pattern classification and knowledge discovery (Yang and Honavar, 1998). It was also used with neural networks for selecting features for defect classification of wood boards (Caballero and Estevez, 1998). Hong and Cho (2006) proposed GA with neural network to select feature subset to get high accuracy for classification. Similarly GA with neural network was proposed for feature selection for the classification of different types of small breast abnormalities (Zhang *et al*., 2004). Another ML algorithm that was employed as a fitness function of GA is support vector machine. To cite examples, Eads *et al*. (2002) used GA with support vector machine for feature selection in time series classification. Frohlich (2004) investigated GA with support vector machine and compared them with other existing algorithms for feature selection. Also, Morariu *et al*. (2006) presented GA with a fitness function based on the support vector machine for feature selection which has proven to be efficient for nonlinearly separable input data. For the classification of hyper-spectral data, GA with support vector machine was proposed by Zhuo *et al*. (2008). Yu and Cho (2003) proposed a feature selection approach, in which GA was employed to implement a r andomized search and support vector machine was employed as a base learner for keystroke dynamics identity verification.

GA with decision tree, (e.g., ID3, C4.5) was explored in (Vafaie and De Jong, 1995; 1992) to find the best feature set to be used by the induction system on difficult texture classification problems. William (2004) designed a generic fitness function for validation of input specification and then used it to develop GA wrapper for feature selection for decision tree inducers. The effectiveness of GA for feature selection in the automatic text summarization task was investigated by Silla *et al*. (2004) where the decision tree was used as a fitness function.

In the filter mode, as a subset search approach, GA seems more computationally attractive than in the wrapper mode. This is because the computational time of GA tends to be high and the run of the ML algorithm is needed every time a chromosome in GA population is evaluated. Therefore combining it with the ML algorithm in a wrapper mode is not so efficient. Some examples of using GA in the filter mode include (Liu *et al*.,1995), in which mutual information measurement between classes and features were used as evaluation function. Based on the experimental results of h and written digit recognition, this method reduces the number of features needed in the recognition process without impairing the performance of the classifier significantly. Ozdemir *et al*. (2001) used GA for feature selection by minimizing a cost function derived from the correlation matrix between the features and the activity of interest that is being modeled. In this work, from a dataset with 160 features, GA selected a feature subset (40 features) which built a better predictive model than with full feature set. Another example is the work of Zhang and Hu (2005), in which GA was used with Mutual Information (MI) to evolve a near optimal input feature subset for neural networks. A fast filter GA approach for feature selection which improves previous results presented in the literature of feature selection was described by Lanzi (1997).

As a hybrid approach for feature selection (Shahla *et al*., 2009), GA was investigated by Cantu-Paz (2004), in which GA and a method based on class separability applied to the selection of feature subsets for classification problems. This approach is able to find compact feature subsets that give the most accurate results, while beating the execution time of some wrappers. A feature selection approach named Relief F-GA-Wrapper was proposed by Zhang *et al*. (2003) to combine the advantages of filter and wrapper. In this approach, the original features are evaluated by the ReliefF filter approach and the resulting estimation is embedded into the GA to search optimal feature subset with the train accuracy of ML algorithm for h and written Chinese characters dataset. Additionally, Fatourechi *et al*. (2007), proposed two stages feature selection. The first stage employs mutual information to filter out the least discriminate features, resulting in a reduced feature space. Then a GA is applied to the reduced feature space to further reduce its dimensionality and select the best set of features.

In addition to the aforementioned applications, GA continues to attract researchers to combine it with others techniques to improve the efficiency of feature selection in various ways (Shahla *et al*., 2009; Yang *et al*., 2011). Gheyas and Smith (2010) have improved a version of GA that tackles feature selection problem.

An important aspect of the previous applications of GA for feature selection is that the standard GA with fixed length binary representation scheme to represent each chromosome of the population as a feature subset. For an n-dimension feature space, each chromosome is encoded by an n-bit binary string b1 ... bn. $b_i = 1$, if the ith feature is present in the feature subset represented by the chromosome and $b_i = 0$ otherwise. Figure 4 is a hypothetical chromosome represented using the fixed length binary representation scheme of the standard conventional GA.

The advantage of this representation is that the standard GA can be used straightforward without any modification. Unfortunately, the fixed length binary representation is appropriate if the dimension of the data is not high. As the dimension of the data becomes huge, the chromosome becomes very long and the evolution of GA becomes inefficient (Arbor *et al*., 2006).The case even worsens when only small number of these features is needed. There have been several attempts to tackle this problem and apply GA for feature selection in high dimensional data (Liu and Yu, 2005; Sanchez-Ferrero and Arribas, 2007; Lecocke and Hess, 2007; Arbor *et al*., 2006; Silla *et al*., 2004). These attempts are based on a simplification assumption of pre-specifying the number of the features that must be selected.

Accordingly, the chromosome encodes the indices of the selected features, rather than the presence or absence of each feature. Figure 5 depicts the chromosome representation adopted by these works.

Although the above representation facilitates the application of the standard GA, the assumption of having pre-specified number of features is not correct and need a prior knowledge about the domain to estimate the number of features that must be selected. Alternatively, this study presents an efficient solution to the problem that exploits the variable length representation scheme of GA and encodes each chromosome as the selected subset of feature. However, using variable length representation calls for modifying the genetic operators or devising new operators to cope with the new representation scheme. In the following sections, the elements of the alternative VLGA developed for feature selection are described in details.

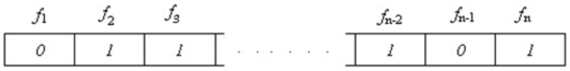


Fig. 4: Fixed length chromosomes for feature selection

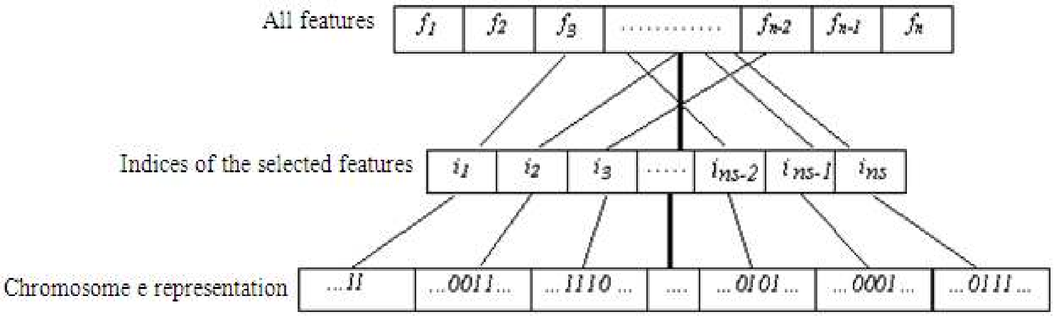


Fig. 5: Chromosome representation for high dimensional data

MATERIALS AND METHODS

VLGA for feature selection in high dimensional data: we describe the proposed approach for feature selection in high dimensional data. The proposed approach is essentially a variable length GA developed specifically for this task. Before diving into the details of the proposed approach, it is worth mentioning that the idea of using variable length representation in the context of evolutionary algorithms is as old as the algorithms themselves. Fogel and Walsh (1966) seem to be among the first experimenting with the variable length representation. In their work, they evolved finite state machines of a varying number of states, therefore making use of operators like addition and deletion. Holland and Holland (1975) proposed the concepts of gene duplication and gene deletion in order to raise the computational power of evolutionary algorithms. Smith departed from the early fixed-length character strings by introducing variable length strings, including strings whose elements were if-then rules (rather than single characters) (Smith, 1980). Since the first attempts of using variable length representations, many researchers made use of the idea under different motivations such as engineering applications (Davidor, 1991a; 1991b) or raising the computational power of evolutionary algorithms (Schtz, 1997).

With regard to GA as an evolutionary algorithm, the use of variable length representation has been proposed in several versions. Well known versions are messy GA (Goldberg *et al*., 1990), genetic programming (Koza, 1992) and species adaptation GA (Harvey, 1995). These versions differ in the specification of the representation scheme and consequently the genetic operators. Messy GA uses binary representation in which each gene is represented by a pair of numbers that are the gene position and the gene value. Messy GA uses the mutation operator as with standard GA. Instead of crossover, messy GA uses the splice and cut operators. The splice operator joins one chromosome to the end of the other. The cut operator splits one chromosome into two smaller chromosomes. Genetic programming is an extension of GA with variable length representation scheme in the form of hierarchical tree representing computer program. The aim of genetic programming is to find the best tree (computer program) that solves a given problem. It adapts genetic operators of the standard GA to cope with the tree representation scheme. Species adaptation GA uses a variable length binary representation scheme. It differs from the standard GAs subtly but significant. Evolution is directed by selection exploiting differences in fitness causes by variations in the genetic makeup of the population. While mutation operator in the standard GA and genetic programming is considered as a background operator and crossover is usually assumed to be the primary operator, in species adaptation GA the reverse is true. Of these two genetic operators, mutation is primary and crossover, though useful, is secondary. Besides that, researchers may opt to develop a domain-specific version of variable length representation GA to better meet the requirements of the domain, rather than using existing ones. For example, Zebulum *et al*. (2000) investigated the application of GA in the field of evolutionary electronics, in which a special variable length GA was proposed to cope with the main issues of variable length evolutionary systems. Following this trend, in this study, a special version of GA for feature selection in domains with huge dimensional data is developed as described below.

Representation scheme: The representation scheme of the proposed VLGA approach is based on variable length non-binary representation scheme, in which each chromosome represents the selected subset of features. It is a direct representation scheme with no encoding or decoding process to map between the genotype and the phenotype levels. Figure 6 is a hypothetical chromosome represented using this scheme. An interesting aspect of this representation scheme is that it is positional independent meaning that the gene position has no role in determining the aspects of the chromosome at the phenotype level.

Feature space mask: Technically speaking, GA explores the promising points in the search space via genetic operations, therefore, the representation scheme and the genetic operators should give rise to an effective exploration of the search space. Using the proposed representation scheme directly does not assist the genetic operators to explore new points in the search space. Therefore, to ensure a good exploration of the feature space, we propose a feature space mask. The feature space mask is a binary string with length equal to the size of feature space, in which each bit marks the status of a single feature in the feature space.

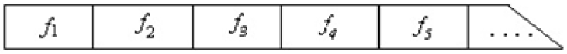


Fig. 6: Variable length chromosome for feature selection

Accordingly, the value 1 indicates that the feature is being used by the current population and the value 0 indicates that the feature is not in use. Figure 7 shows feature space mask schematically. It shows that the feature f_2, f_3 and f_n are participating in the current GA population, whereas the feature f1, f_{n-2}, f_{n-1} are not. As it will be described, the status of the feature in the feature space mask is updated either immediately after performing a genetic operator or through a rebuilding step of the phrase space mask which takes place during the transition from generation t to generation t + 1.

Fitness function: The fitness function is the driving force for the evolution in GA. For feature selection it is the evaluation function that evaluates a candidate subset of features. As the aim of feature selection is to find a minimum number of features with a maximum in formativeness, the fitness function of a subset p consists of a combination of two measures, namely subset in formativeness and subset complexity as follows:

$$f(p) = \text{Info}(p) - pf * \frac{L(p)}{N} \quad (1)$$

In this formula, Info (p) denotes the estimated in formativeness of the features subset p and L (p) is a measure for the complexity of the feature subset usually the number of utilized features. Furthermore, N is the feature space cardinality and pf is a punishment factor to weigh the multiple objectives of the fitness function. The number of features used by a subset is intended to lead the algorithm to regions of small complexity.

VLGA selection scheme: The selection scheme of the proposed VLGA approach for feature selection is (k, q) tournament selection. It randomly chooses k chromosomes from the current population and with certain probability q returns the best chromosome, otherwise return the worst chromosome.

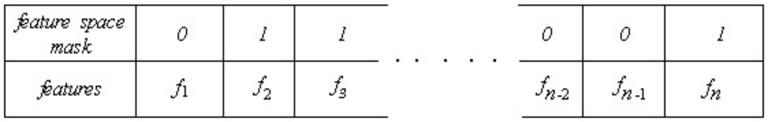
VLGA genetic operators: The proposed VLGA approach makes use of three genetic operators modified from the standard GA to cope with the variable length representation scheme. Furthermore, it introduces a new operator called AlterLength.

Reproduction: The reproduction operator of the proposed VLGA is similar to the reproduction operator of standard GA. With the reproduction probability Pr, a chromosome is randomly selected from the current generation and then copied into the new generation without any modification.

Crossover: To cope with the variable length representation of the proposed VLGA approach, the uniform crossover of the standard GA has been adapted. The uniform crossover (Mitchell, 1996) is an operator that decides with a probability which parent will contribute to each of the gene values in the offspring chromosomes.

For the proposed VLGA, the uniform crossover has been modified as follows. First two parents (chromosomes) from the current population are selected. Then with a probability 0.5 the length of the offspring is chosen to be either the length of the short or long parent. If the length of the short parent is chosen, then a uniform crossover is performed between the short parent and an equal length segment from the long parent. If the length of the parent is chosen, then a uniform crossover is performed between the short parent and an equal length segment of the longer parent. The remaining parts of the long parent are appended to the beginning and the end of the offspring. Figure 8 shows VLGA uniform crossover schematically.

Mutation: The proposed approach for mutation is to replace the values of some genes by new values from the feature space which are not participating in the current population. The mutation operator is applied with probability P_m to each chromosome generated from the crossover operation. This operator is performed with assistance of the feature space mask. More specifically, for each gene in the chromosome, if it is selected for mutation then it is replaced by a r randomly selected feature from the feature space which has its status marked inactive and then the status of the selected feature in the features space mask is set to active immediately. With regard to the mutated feature (gene), its status in the feature space is not set to inactive immediately because this feature is still in use by the parents (members of the current GA population). Setting the status of the mutated gene to inactive is performed after all genetic operations on the current population are completed and a rebuilding step for the feature space mask is performed. Figure 9 shows mutation operator schematically.



feature space mask	0	1	1	. . . . .	0	0	1
features	f_1	f_2	f_3		f_{n-2}	f_{n-1}	f_n

Fig. 7: Feature space mask

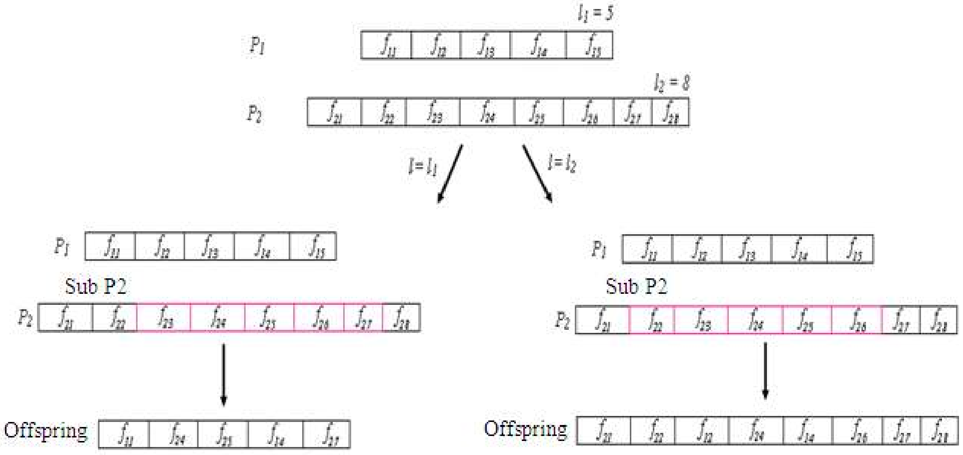

$l_1 = 5$
P_1
f_{11} f_{12} f_{13} f_{14} f_{15}
$l_2 = 8$
P_2
f_{21} f_{22} f_{23} f_{24} f_{25} f_{26} f_{27} f_{28}
$l = l_1$
$l = l_2$
Sub P2
Offspring f_{11} f_{24} f_{25} f_{14} f_{27}
Offspring f_{21} f_{22} f_{12} f_{24} f_{14} f_{26} f_{27} f_{28}

Fig. 8: VLGA uniform crossover operator

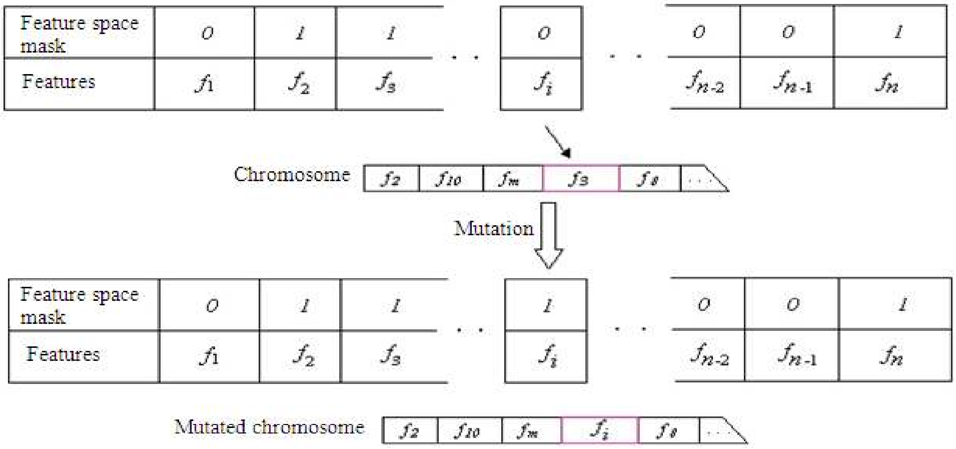

Feature space mask
Features
0 1 1 . . 0 . . 0 0 1
f_1 f_2 f_3 f_i f_{n-2} f_{n-1} f_n
Chromosome f_2 f_{10} f_m f_3 f_8 . . .
Mutation
0 1 1 . . 1 . . 0 0 1
Mutated chromosome f_2 f_{10} f_m f_i f_8 . . .

Fig. 9: Example of *VLGA* mutation operator

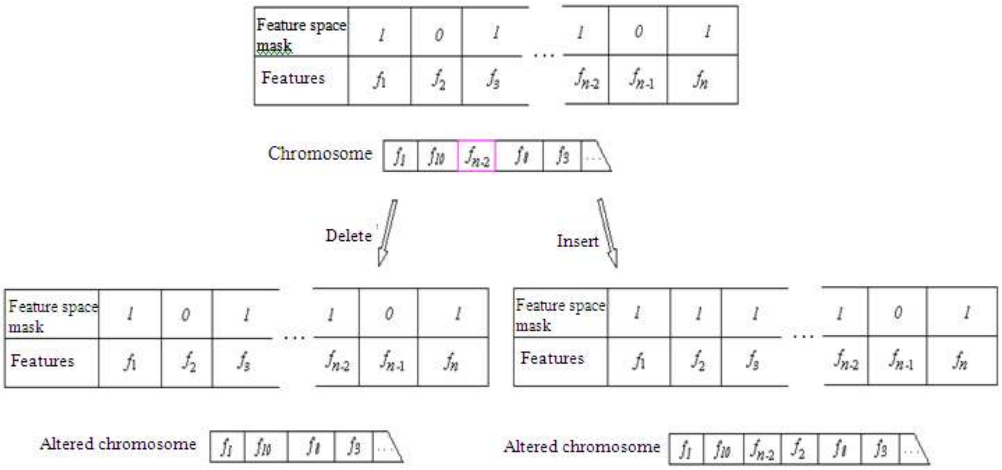
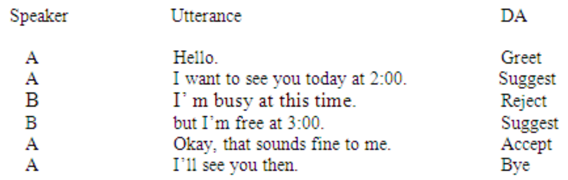

Feature space mask
Features
1 0 1 . . . 1 0 1
f_1 f_2 f_3 f_{n-2} f_{n-1} f_n
Chromosome f_1 f_{10} f_{n-2} f_8 f_3 . . .
Delete
Insert
1 0 1 . . . 1 0 1
1 1 1 . . . 1 0 1
Altered chromosome f_1 f_{10} f_8 f_3 . . .
Altered chromosome f_1 f_{10} f_{n-2} f_2 f_8 f_3 . . .

Fig. 10: Example of AlterLength operator



Speaker	Utterance	DA
A	Hello.	Greet
A	I want to see you today at 2:00.	Suggest
B	I' m busy at this time.	Reject
B	but I'm free at 3:00.	Suggest
A	Okay, that sounds fine to me.	Accept
A	I'll see you then.	Bye

Fig. 11: Hypothetical dialogue annotated with DAs

Alter length: The crossover and mutation operators are designed specifically to introduce variation to the content of the chromosome. To introduce a variation to the length of the chromosome, the alter length operator is proposed. The alter length operator r randomly exp ands (shrinks) the chromosome by inserting (deleting) a single feature to (from) the chromosome. In case of insertion, the inserted feature is r anomaly selected from inactive features in the feature space. In case of deletion the selected feature is deleted from the chromosome and its status in the feature space mask remains active until the rebuilding step of the feature space mask. The AlertLength operator is performed with a probability Pal. as shown in Fig. 10.

Stopping criterion and control parameters: The proposed VLGA approach makes use of the standard stopping criteria used in the conventional GA which are either to stop after a pre-defined number of generations or to stop when the evolution does not introduce any significant evolution. Regarding the control parameters, VLGA uses the following parameters: Population Size (PopSize), tournament selection parameters (q, k), reproduction probability (Pr), crossover probability (Pc), mutation probability (Pm) and alter length probability (Pal).

Case study:

VLGA for lexical cue selection: To evaluate the proposed VLGA approach for feature selection in huge dimensional data, it has been applied for feature selection in the context of designing dialogue act recognition (DAR) model. To underst and the context of the VLGA application, we start with a brief description of the lexical cue selection in the context of DAR.

Lexical cue selection for DAR: Dialogue Act (DA) is defined as a concise abstraction of a speaker's intention in his utterance. It has roots in several language theories of meaning, particularly speech act theory (Austin, 1962), which interprets any utterance as a kind of action, called speech act and categorizes speech acts into speech acts categories (Searle, 1975). DA, however, extends speech act by taking into account the context of the utterance (Bunt, 1994). Figure 11 is a hypothetical dialogue annotated with DAs.

The automatic recognition of DA, Dialogue Act Recognition (DAR), is a task of crucial importance for the processing of natural language at discourse level. It is defined as follows: Given an utterance with its preceding context, how to determine the DA it realizes. Formally, it is a classification task in which the goal is to assign a suitable DA to the given utterance. Due to its importance for various applications such as dialogue systems, machine translation, speech recognition and meeting summarization, it has received a considerable amount of attention (Jurafsky, 2004).

Recently, Machine Learning (ML) techniques have become the current trend for tackling the DAR problem (Fishel, 2007). In this regard, various ML techniques have been investigated and the resulting models have become known as cue-based models (Sridhar *et al*., 2009; Keizer and Akker, 2007). ML technique builds a cue-based model of DAR by learning from utterances of a dialogue corpus the association rules between surface linguistic features of utterances and the set of DAs. In doing so, ML exploits various types of linguistic features such as cue phrases, syntactic features, prosodic features. Among different types of linguistic features, cue phrases are the strongest (Jurafsky *et al*., 1998). They are defined by Hirschberg and Litman (1993) as linguistic expressions that function as explicit indicators of the structure of a discourse. Since not all phrases are relevant to the DAR, prior to applying a ML technique the selection of relevant cue phrases is of crucial importance. A successful selection of cue phrases would speed up the learning process, reduce the required training data and improve the classification accuracy (Blum and Langley, 1997).

One cue-based model, which has been used as a context of the current research, is Dynamic Bayesian Network (DBN) model (Yahya *et al*., 2006; 2009). As depicted in Fig. 12, the DBN model of DAR consists of T time slices, in which each slice is a Bayesian Network (BN) composed of a number of r random variables. The DBN models a sequence of utterances over time in such a way that each BN corresponds to a single utterance. In this sense DBN is time invariant, meaning that the structure and parameters of BN is the same for all time slices. Moreover, in each BN, there is a hidden random variable which represents the DA that need to be recognized and a set of observation variables extracted from the linguistic features of the corresponding utterance. In this model, dynamic Bayesian ML algorithms have been employed to construct the DBN model from a dialogue corpus.

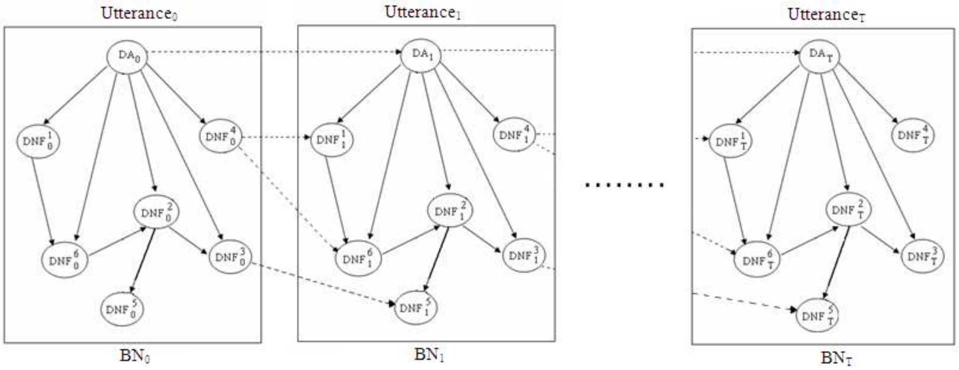


Fig. 12: Example of the DBN model of DAR

Table 1: Examples of ranking approaches metrics

Metric	Formula
MI	$\mathrm{MI}(f,c)=\log\frac{P(f,c)}{P(f).P(C)}$
OR	$\mathrm{or}(f,c)-\frac{P(f\|c).(1-P(f\|\overline{c}))}{(1-P(f\|\overline{c}).P(f\|c))}$
IG	$\mathrm{IG}(f,c)=\sum_{c\in\{c,\overline{c}\}}\sum_{f\in\{f,\overline{f}\}}p(f,c)\log\frac{P(f,c)}{P(f).P(c)}$
χ^2	$x(f,c)=\frac{N[P(f,c).P(\overline{f},\overline{c})-P(f,c).P(\overline{f},c)]^2}{P(f).P(\overline{f}).P(c).P(\overline{c})}$

An essential issue aroused while building the DBN model or DAR is the specification of the observation variables. For this model, it has been suggested that the number of the random variables in each BN should be equal to the number of DAs that the model recognizes. Moreover, each variable is defined as a logical rule, Disjunctive Normal Form (DNF), consists of a set of cue phrases which are informative to one and only one DA and expressed as follows:

$$\mathrm{DNF} = \text{if } (p_1 \equiv p_2 \equiv \ldots \equiv p_m) \qquad \text{then} \qquad \mathrm{DA}$$

where, DI is the target DI and pi is a cue phrase selected for that DA. In doing so, each variable works as a binary classifier for the given DA:

$$\mathrm{Info}(p)=\frac{TP+TN}{NU} \tag{2}$$

Where:

TP = The number of time the selected phrases give true when the utterance belongs to the target DA
TN = The number of times the selected phrases gives false when the utterance does not belong to the target DA
NI = The total number of utterances in the dialogue corpus.

As the literature indicates that only the ranking approaches have been investigated (Samuel *et al*., 1999; Webb *et al*., 2205; Lesch, 2005; Kats, 2006; Verbree *et al*., 2006) due to their computational efficiency, regardless of their inefficiency with respect to the relevance and redundancy of the selection. In addition to examining the proposed VLGA on the lexical cue selection, a number of ranking approaches which are always applied for cue phrase selection have been experimented. The overall procedure of these approaches is to score each potential feature according to a particular metric and then pick out the best n features. Table 1 contains a list of the ranking approaches that have been selected as a baseline approaches. In their formulas f denotes the feature and c denotes the class which represent a phrase and a DA respectively.

Settings of the experiments: To experiment the proposed VLGA and the baseline approaches on lexical cues selection in the abovementioned context, SCHISMA dialogue corpus (Andernach *et al*., 1995), a collection of 64 dialogues in the domain of information exchange and transaction in a theater, has been annotated with DAs from DAMSL coding scheme (Allen and Core, 1997). First each utterance is subdivided into one or more segments and the dialogue acts are assigned to the segments. In the current study, we focus on the following Das:

- After annotating the corpus with DAs, the following processes were performed to generate the phrases space
- Tokenization: Tokenization occurs at utterance level and the token is defined as a sequence of letters or digits separated by separator (e.g. ,”.” , “:” , “;”). In this process, all punctuations are discarded except “?” which is treated as token
- Removing morphological variations: It has been noticed that most of morphological variations in SCHISMA corpus are the plurals and tenses

variations which are not significant for the recognition process

- Semantic clustering: Clusters certain words into semantic classes based on their semantic relatedness and then replace each occurrence of the words with the cluster name. For SCHISMA corpus, the following semantic clusters were identified
  - Show Name: Any show name appears in the corpus
  - Player Name: Any player name appears in the corpus
  - Number: Any sequence of digits (0. . . 9) or alphabetic numbers(one, two, . . .)
  - Day: Any occurrence of a day name (Monday, Tuesday, . . . , Sunday)
  - Month: Any occurrence of a month name (January, December)
  - Price: A Number cluster preceded by currency symbol (f , ff)
  - Date: Any of the following sequences <Number Month Number>, <Month> <Number>, <Number Month>
  - Time: A Number cluster preceded by the proposition “at”

N-gram phrases generation: In this process all phrases that consist of one, two and three words were generated from each utterance in the corpus.

Removing less frequent phrase: To reduce the dimension of the phrases space, the phrases occur less than a frequency threshold number were removed. Based on the experiments of Webb *et al*. (2005), the frequency threshold was 3.

The above preprocessing steps resulted in a phrases space of 1336 phrase. This phrases space was used in the experiments of the baseline approach. However, in the subsequent experiments further preprocessing steps were introduced which make the phrases space for each DA has different size.

RESULTS

In the following, the results obtained from four experimental cases are presented.

Baseline approaches: Each of the ranking approaches listed in Table 2, was experimented on the selection of cue phrases for each DA. More specifically, for each DA, each ranking approach ranked the phrases using its own metric. Then, the fitness value, F(p), along with in formativeness value, Info(p) and complexity value, L(p)/N of each k phrases (k = 1, 2, ... n) in the ranked list were calculated and the top k phrases that maximize the fitness value, F(p), is the selected set of phrases for that DA.

VLGA Approach: Case 1: The aim of this case of experiments is to evaluate the proposed VLGA approach on the selection of cues phrases for each DA given in Table 6. The settings of the control parameters are as follows PopSize = 500, q = 10, k = 0.7, r = 0.3, Pc = 0.7, Pr = 0.1, Pal = 0.2, Pm = 0.1 and the stopping criterion is to stop if there is no significant improvement within 10 generations. Table 5 summarizes the results obtained from this case of experiments. It should be mentioned that the selection of the parameter’s values was in light of (Mitchell, 1996) and some sensitivity experiments for some parameters such as PopSize, Pc, Pr, Pm, Pal have been performed and the best values found have been selected. Additionally, the run of VLGA was repeated five times and the results of the best run were reported.

Figure 13-14 are example of the GA evolution during the selection of the cue phrases for statement DA. The curves correspond to the best evolutionary trends. In general, it can be noticed that there is a rapid growth at the early generations followed by a long period of slow evolution until meeting the stopping criterion. This reflects the nature of the search space of cue phrases which is hugely multimodal and contains a lot of peaks. An interesting aspect of the average population fitness curve is that despite the fluctuations, an overall look at the curve shows a general tendency to improving the average fitness value, particularly at the early generations.

Table 2: Experimented DA and their frequencies in SCHISMA corpus

DA	Meaning	Frequency
Statement	The speaker makes a claim about the world	817
Query-if	The speaker asks the hearer whether something is the case or not	108
Query-ref	The speaker asks the hearer for information in the form of references that satisfy some specification given by the speaker	598
Positive-answer	The speaker answer in positive	561
Negative-answer	The speaker answer in negative	72
No-Blf	The utterance is not tagged with any blf DA	968

Table 3: Results of ranking approaches experiments

	MI			IG			χ2			OR		
	Info(P)	L(P)/N	F(P)	Info(P)	L(P)/N	F(P)	Info(P)	L(P)/N	F(P)	Info(P)	L(P)/N	F(P)
Statement	0.8403	0.1055	0.7347	0.6385	0.0052	0.6385	0.6619	0.0007	0.6612	0.7675	0.0509	0.7166
Query-if	0.9634	0.0045	0.9589	0.9580	0.0007	0.9580	0.9580	0.0007	0.9572	0.9599	0.0030	0.9569
Query-ref	0.8505	0.0404	0.8101	0.8149	0.0060	0.8149	0.8149	0.0060	0.8089	0.8803	0.0412	0.8391
Positive-answer	0.8217	0.0636	0.7581	0.7333	0.0015	0.7333	0.7860	0.0007	0.7853	0.8105	0.0464	0.7640
Negative-answer	0.9687	0.0015	0.9672	0.9595	0.0015	0.9595	0.9595	0.0015	0.9580	0.9687	0.0022	0.9665
No-blf	0.8036	0.1198	0.6839	0.7333	0.0015	0.7333	0.7333	0.0015	0.7318	0.7851	0.0786	0.7065

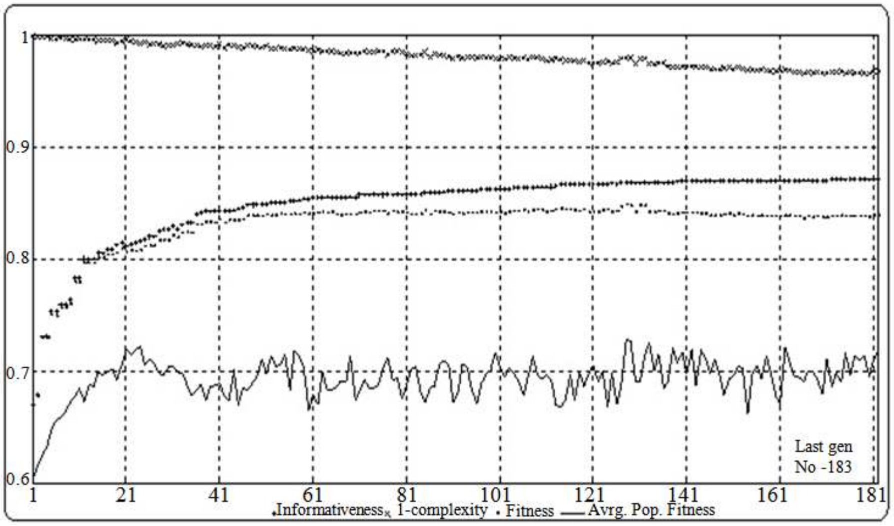


Fig. 13: GA evolution of cue phrases selection for statement DA

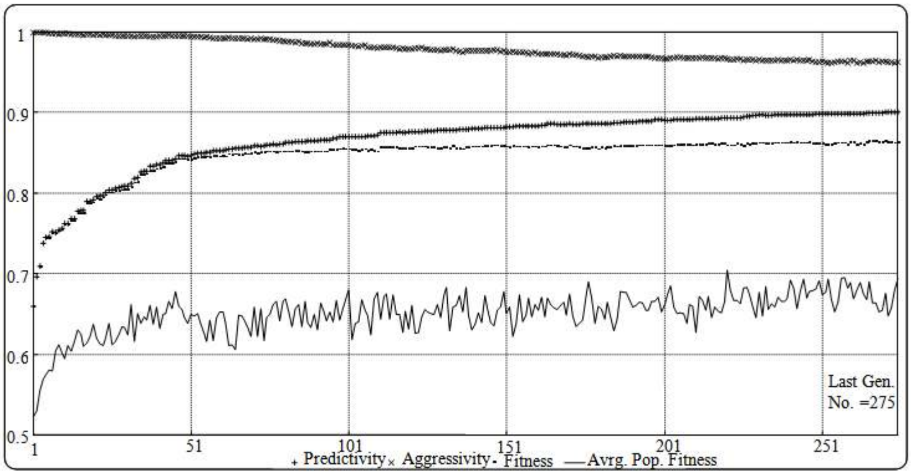


Fig. 14: GA evolution for cue phrases selection for statement DA

VLGA approach: Case 2: The aim of this case of experiment is to test the scalability of the proposed VLGA approach on the selection of lexical cue from a larger space. To do so, first the cardinality of the phrase space been increased by introducing two source of information to the phrases, that are the type of lexical cue and it position within the utterance. To understand the characteristics of the new space and the feature selection, each phrase has two sides, positive and negative and accordingly the phrase is classified either positive or negative depending on the dominant side. An efficient way to exploit negative phrases, which was described by Zheng *et al*. (2007) is to select positive and negative phrases independently based on their use. For cue phrase selection in DAR, the positive phrases can be used to indicate the membership of an instance to the target DA and the negative phrases can be used to help in increasing the relevancy of the positive phrases by confidently rejecting instances which do not belong to the target DA but still contain the positive phrases. For example, in SCHISMA dialogue corpus, the positive phrase "ticket" is relevant for both the statement and query-ref DAs. To increase the relevancy of this cue phrase for the statement DA, negative cue phrase such as "how much" and "?", which are relevant to the query-ref DA, yet not to the statement DA, might be selected and conjuncted with the "ticket" to accept only the utterances that contain "ticket" and does not contain "how much" and "?". In general the aim here is to select positive and negative cue phrases that meet the following expression where ppj is a positive phrase and npj ... npk j are negative phrases associated with it:

$$\begin{aligned}&\text{if} \quad ((pp_1 \wedge !np_1 \ldots \wedge !np_{k1}) \\ &\vee \ldots \vee (pp_j \wedge !np_j \ldots \wedge !np_{kj}) \vee \ldots \vee \\ &(pp_m \wedge !np_m \ldots \wedge !np_{km})) \quad \text{then} \quad DA\end{aligned}$$

To account for the negative cue phrases, each phrase occur within the utterance that belongs to the target DA is marked positive and each phrase occur within the utterance that does not belong to the target DA is marked negative. It could happen that some phrases occur in utterances that are labeled with the target DA and in utterances not labeled with the target DA, hence it might be possible to find tow identical phrases marked with negative and positive. Table 5 summarizes the results obtained from this stage of experiment.

The information about phrase's position within an utterance is useful to increase its relevancy for a given DA. We conducted this stage of experiment to investigate the ability of the genetic-based approach to select cue phrases after incorporation of the phrase's positional information. To do that, each positive phrase in the phrase space was marked with one of three possible positional labels, which represent the position of the phrase within the utterance. These labels are: Begin, if the phrase occurs at the beginning of any utterance labeled with the target DA, End, if the phrase occurs at the end of any utterance labeled with the target DA and Contain, if the phrase occurs elsewhere. It might happen that certain phrase occurs in different positions within the utterance. In this case, multiple instances of this phrase, each with different position label, are created. Consequently, each DA has a different size search space as shown in Table 5. The genetic-based approach was applied with the same parameters specified in the previous stages and the results are shown in Table 6.

Validation experiments: The aim of this stage of experiments is to validate the use of the proposed genetic-based approach for the ML algorithm application. More specifically, the cues phrases generated from the above experiments were used to build DBN model for DAR. The hypothesis is that the more relevant cue phrases, the more accurate DAR. First the sets of cue phrases generated by the genetic-based approach in each of the previous stages were used to specify the DBN random variables as described in earlier, so that each random variable is a binary classifier for a single DA. Then DBNs ML algorithms were used with 10-fold cross-validation to construct the structure of the DBN model, assess its parameter and estimate its recognition accuracy using probabilistic networks library (Intel, 2004) which is freely available from http://www.intel.com/research/mrl/pnl . The same experiment was repeated using the sets of cue phrases generated by MI and the results of these experiments are summarized in Table 6.

Table 4: Results of genetic-based approach experiments

DA	Info (p)	L(p)/N	F(P)
Statement	0.871031	0.032186	0.838845
Query-if	0.964338	0.003743	0.960595
Query-ref	0.903273	0.005988	0.897285
Positive-answer	0.852467	0.019461	0.833006
Negative-answer	0.96922326	0.001497	0.96772626
No-blf	0.81143135	0.016467	0.79496435

Table 5: Results of genetic-based approach with cue positional information experiments

DA	Phrases search space	Relev (p)	L(p)/N	F(P)
Statement	2329	0.900342	0.036926	0.863416
Query-if	1625	0.980459	0.008615	0.971844
Query-ref	2047	0.926234	0.025892	0.900342
Positive-answer	2377	0.90083051	0.036601	0.86422951
Negative-answer	1581	0.98876405	0.008855	0.97990905
No-blf	2354	0.84465069	0.029737	0.81491369

Table 6: Accuracies of the DAR models

	Recognition accuracy		
Selection approach	Min.	Max.	Avrg.
Ranking approach (MI)	76.74	78.48	77.72
VLGA Case 1	76.89	79.19	78.34
VLGA Case 2	77.61	81.09	79.67

DISCUSSION

In the following, the results obtained from four experimental cases are discussed.

Baseline approaches: The results of the baseline approach experiments are given in Table 3. From these results, it can be observed that there is a similarity between the performance of MI and OR from one side and the performance of IG and χ^2 from the other side in three aspects. First, from the complexity values, L(P)/N, it is clear that MI and OR tends to select larger number of phrases than IG and χ^2. Second, the In f o(P) values of MI and OR are higher than IG and χ^2. Third, as a direct result of the similarity in Info(P) and L(P)/N values within each group ,(MI, OR) and (IG, χ^2), the pattern of the fitness values is similar within each group, though, between the two groups the comparison of fitness values are not conclusive.

The similarity between the two groups, (MI, OR) and (IG, χ^2), can be understood through the following facts. For each DA, each phrase has two sides, positive and negative. The positive side depends on the presence of the phrase in the utterances labeled with the target DA and the absence of the phrase from the utterances labeled with other DAs.

The negative side depends on the absence of the phrase from the utterances labeled with the target DA and the presence of the phrase in the utterances labeled with other DAs. Based on that, the ranking approach is classified as either one-sided metric or two-sided metric depending on whether it's metric account for the negative side of the phrase or not (Bunt, 1994) are phrases with the highest positive sides and, definitely, the lowest negative sides. With regard to the two-sided metrics, they rank the phrases according to a combination of both positive and negative sides. Therefore the top k phrases in the list are phrases with the highest negative or positive sides.

From Table 2, it is clear that MI and OR are one-sided metrics and IG and χ^2 are two-sided metrics. It is also obvious that the fitness measure, Eq. 2, which was used for the selection of cue phrases from the ranked list, has its Info (P) subpart depends on the positive side of the phrases rather than negative side. Therefore, the ranking of the one-sided metrics is more appropriate for the fitness measure than the two-sided metrics which can interpret the higher Info(P) values of the cue phrases selected by MI and OR. However, the inability of these approaches to account for the correlation between cue phrases lead to the selection of large number of cues phrases in case of MI and OR. In other words, the ranking approaches assume that the relevance of a set of phrases is equal to the summation of the individual relevance of each phrase which leads to redundant selection.

The general conclusion that can be drawn from this stage of experiments is that the ranking approaches are not able to maintain a tradeoff between the two subparts of the fitness functions. They tend to optimize one subpart at the expense of the other.

VLGA approach: Case 1: The results of this case of experiments are empirical evidences on the efficiency of the VLGA approach for the selection of useful features from huge data. More detailed a comparative look at the result of Table 3-5 shows that the VLGA approach outperforms the ranking approaches for cue phrase selection. This is obvious from the differences between the fitness values, F (P), for VLGA and ranking approaches. The informativeness values, Info(P), of the VLGA approach are higher than their corresponding values of the ranking approaches. With regards to the complexity of the selected cue phrases, L (p)/N, it is obvious that the VLGA approach tends to select smaller number of phrases than MI and OR ranking approaches, yet more than IG and x^2 to confirm that, a paired t-test of the statistical significance of the difference between the F(p) values for both MI ranking approach and VLGA approach was performed at level P < 0.05 and 5 degree of freedom. The obtained t value (t = 3.1123) shows that the difference is statistically significant.

The above findings are direct results of the ability of the VLGA approach to maintain a trade off in formativeness for complexity which can be attributed to two factors. First, in the VLGA approach, the evaluation and the selection processes are based on the fitness measure which depends on the subsequent use of the selected phrases. In contrast to that, in the ranking approach the evaluation of the phrases is based on the ranking approach metric which evaluate the phrase based on the intrinsic properties of the phrases whereas the selection depends on the fitness measure. The second factor is the ability of the VLGA approach to account for the correlation between the selected cue phrases. Unlike the ranking approaches, the VLGA approach evaluates the selected cue phrases as whole rather than evaluating each phrase individually and then assuming the relevancy of the set is equal to the summation of the individual relevancy of each phrase which leads to redundant selection.

VLGA approach: Case 2: It appears from the results in Table 6 that there is an improvement in the fitness values, F(P), of the selected cues for each DA which can be attributed to the improvement of the corresponding Relev(P) values due to using the cue's positional information. In terms of complexity, L(P)/N, there is a slight decrease in its values for some DAs, however there are cases, where the L(P)/N values are similar or even better, for instance in positive answer DA, there is an obvious improvement in both values, Relev(P) and L(P)/N. This is empirical evidence on the ability of the genetic-based approach and on the role of the positional information.

The analysis of the statistical significance of the difference between the fitness values, F(P), of Table 6 and Table 5 using paired t-test at level $p < 0.05$ and 5 degree of freedom confirm this conclusion. The obtained t value ($t = 4.0410$) shows that the difference is very statistically significant.

Validation experiments: It is clear that the difference between the performances of the genetic base approach and MI in cue phrases selection affect the accuracy of the DBNs of DAR on the basis that the better cues selections approach, the higher recognition accuracy. To underst and the influence of the cues phrase selection approaches on the recognition accuracy, it should be borne into mind that the construction of the DBNs models of DAR is based on the binary representation of the datasets which are resulted from the extraction of the random variables from the utterances. In this representation, utterances that belong to a certain DA should have a distinct pattern, which is composed in the ideal case of n -1 bit with 0s values and a single bit with 1 value (n is the number of random variables) that corresponds to the random variables of this DA. It is obvious also that the quality of the representation depends on the relevancy of the selected cues phrases that form the random variables. In other words, the better cues selection approach, the better data representation and consequently the better constructed DBNs models.

CONCLUSION

In this study, an adapted GA approach for feature selection in huge dimensional data is introduced. The proposed approach is a variable length GA with specialized genetic operator developed specifically for this task. Several stages of experiment were conducted and the obtained results suggest a number of important conclusions. Firstly, the results confirm that the ranking approaches are not the optimal approaches for cues phrases selection in DAR and similar high dimensional domains. The selection in these approaches is independent of the subsequent use and they are not able to account for the correlation between the selected features. Secondly, the results of the proposed genetic-based approach shows the ability of the genetic-based approach to account for the correlation between the selected cues enables them to select a minimal number of relevant phrases. It is apparent from the high reduction of the number of the selected cues. Thirdly, In contrast to the ranking approaches, the proposed genetic-based approach shows its ability to exploit the negative phrases to increase the relevancy of the selected cue phrases. Fourthly, the results confirm that the cue's positional information is useful to improve the relevancy of the selected cue phrases. In general the proposed genetic-based approach has proved its efficiency for the selection of useful cue phrases for DAR. Finally, although the genetic-based approach was applied to cue phrase selection, it can be applied for feature selection in any similar high dimensional domains.

ACKNOWLEDGEMENT

This study was supported by the Scientific & Engineering Research Center in Najran University, Kingdom of Saudi Arabia under research number NU 17/10.

REFERENCES

Allen, J. 1997. Draft of DAMSL: Dialog Act Markup in Several Layers. National Institute informatics. http://ci.nii.ac.jp/naid/10022006078/

Almuallim, H. and T.G. Dietterich, 1991. Learning with many irrelevant features. Proceedings of Ninth National Conference on Artificial intelligence, (AI'91), AAAI Press, USA., pp: 547-552. http://portal.acm.org/citation.cfm?id=1865761

Andernach, J.A., H.W.L. Doest, H.J.A. Akker, G.F. Hoeven and J. Schaake *et al*., 1995. Language analysis for dialogue management in a theatre information and booking system. Proceedings of the 5th International Conference on Artificial Intelligence, (ICAI'95), University of Montpellier, Montpellier, France, pp: 351-362. http://eprints.ewi.utwente.nl/9564/

Arbor, A., J.H. Michigan, S.B. Hong and S.B. Cho, 2006. Efficient huge-scale feature selection with speciated genetic algorithm. Patt. Recogn. Lett. 27: 143-150. DOI: 10.1016/J.PATREC.2005.07.009

Austin, J.L., 1962. How to do Things with Words. 1st Edn., Harvard University Press, Boston, pp: 166.

Baldi, P. and S.A. Brunak, 2001. Bioinformatics: The Machine Learning Approach. MIT Press, Cambridge, Mass, ISBN: 10: 9780262025065, pp: 476.

Ben-Bassat, M., 1982. Classification Pattern Recognition and Reduction of Dimensionality. Handbook of statistics 2: Classification, Pattern Recognition and Reduction of Dimensionality, Krishnaiah, P.R. and L.N. Kanal (Eds.). North-Holland Pub. Co., Amsterdam, pp: 773-791. ISBN: 044486217X

Biesiada, J. and W. Duch, 2005. Feature selection for high-dimensional data: A kolmogorov-smirnov correlation-based filter. Comput. Recogn. Syst., 30: 95-103. DOI: 10.1007/3-540-32390-2_9

Bins, J. and B.A. Draper, 2001. Feature selection from huge feature sets. Proceedings of the 8th International Conferences Computer Vision, July, 7-14, IEEE Xplore Press, Vancouver, BC , Canada, pp: 159-165. DOI: 10.1109/ICCV.2001.937619

Blum, A. and P. Langley, 1997. Selection of relevant features and examples in machine learning. Art. Intell. 97: 245-271. DOI: 10.1016/S0004-3702(97)00063-5

Bunt, H., 1994. Context and dialogue control. Think, 3: 19-31. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.1926

Caballero, R.E. and P.A. Estevez, 1998. A niching genetic algorithm for selecting features for neural network classifiers. Proceeding of the 8th International Conference of Artificial Neural Networks, Springer-Verlag, New York, pp: 311-316.

Cantu-Paz, E., 2004. Feature subset selection, class separability and genetic algorithms. Genetic Evol. Comput. 3102: 959-970. DOI: 10.1007/978-3-540-24854-5_96

Das, S., 2001. Filters, wrappers and a boosting-based hybrid for feature selection. Proceedings of the 18th International Conference on Machine Learning, (ICML'01) Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., pp: 74-81. http://portal.acm.org/citation.cfm?id=658297

Dash, M. and H. Liu, 1997. Feature selection for classifications: Intelligent data analysis. Int. J., 1: 131-156. DOI: 10.1016/S1088-467x(97)00008-5

Dash, M. and H. Liu, 2003. Consistency-based search in feature selection. Art. Intell., 151: 155-176. DOI: 10.1016/S0004-3702(03)00079-1

Davidor, Y., 1991a. A Genetic Algorithm Applied to Robot Trajectory Generation. In: Handbook of Genetic Algorithms, Davis, L. (Eds.). Van Nostr and Reinhold, New York, pp: 144-165. ISBN: 0442001738

Davidor, Y., 1991b. Genetic Algorithms and Robotics: A Heuristic Strategy for Optimization. World Scientific, Singapore, Teaneck, NJ., ISBN: 9810202172, pp: 164.

Duch, W., 2006. Filter Methods. In: Feature Extraction, Foundations and Applications, Guyon, I., S. Gunn, M. Nikravesh and L.A. Zadeh (Eds.). Springer, Berlin, Heidelberg, New York, pp: 89-117. ISBN: 10: 3540354875

Eads, D., D. Hill, S. Davis, S.J. Perkins and J. Ma *et al*., 2002. Genetic algorithms and support vector machines for time series classification. Proceedings of 5th Conference on the Application and Science of Neural Networks, Fuzzy Systems and Evolutionary Computation, SPIE, USA., pp: 74-85. DOI: 10.1117/12.453526

Fatourechi, M., G.E. Birch and R.K. Ward, 2007. Application of a hybrid wavelet feature selection method in the design of a self-paced brain interface system. J. Neuro. Eng. Rehabilit., 4: 1-11. DOI: 10.1186/1743-0003-4-11

Fishel, M., 2007. Machine learning techniques in dialogue act recognition. Estonian Science Foundation. http://math.ut.ee/~fishel/doc/publ/darec.ery07.pdf

Fogel, L.J., A.J. Owens and M.J. Walsh, 1966. Artificial Intelligence Through Simulated Evolution. Wiley, New York, pp: 170.

Frohlich, H., O. Chapelle and B. Sch″olkopf, 2004. Feature selection for support vector machines using genetic algorithms. Int. J. Art. Intell. Tools, 13: 791-800. doi:10.1142/S0218213004001818

Gheyas, I.A. and L.S. Smith, 2010. Feature subset selection in large dimensionality domains. Patt. Recogn., 43: 5-13. DOI: 10.1016/J.PATCOG.2009.06.009

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. 1st Edn., Addison-Wesley Public Co., New York, ISBN: 0201157675, pp: 412.

Goldberg, D.E., K. Deb and B. Korb, 1990. Messy genetic algorithms revisited: Studiesin in mixed size and scale. Complex Syst., 4: 415-444. http://www.complex-systems.com/pdf/04-4-4.pdf

Guo, B., R. I. Damper, S. R. Gunn and J. D. B. Nelson, 2008. A fast separability-based feature selection method for high-dimensional remotely-sensed image classification. Patt. Recogn., 41: 1653-1662. DOI: 10.1016/J.PATCOG.2007.11.007

Hall, M., 1999. Correlation-based feature selection for machine learning. PhD Thesis, The University of Waikato. http://www.cs.waikato.ac.nz/~ml/publications/1999/99MH-Thesis.pdf

Hall, M.A., 2000. Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. Proceeding of the 17th International Conferences on Machine Learning, (ICML'00), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., pp: 359-366. http://portal.acm.org/citation.cfm?id=657793

Harvey, I., 1995. The artificial evolution of adaptive behaviour. D. Phil. Thesis, School of Cognitive and Computing Sciences, University of Sussex.

Hastie, T., R. Tibshirani and J.H. Friedman, 2001. The Elements of Statistical Learning. Springer, New York, ISBN: 0387952845, pp: 533.

Haupt, R.L. and S.E. Haupt, 2004. Practical Genetic Algorithms. 2nd Edn., Wiley-IEEE, New York, ISBN: 0471455652, pp: 253.

Hirschberg, J. and D. Litman, 1993. Empirical studies on the disambiguation of cue phrases. Comput. Linguist., 19: 501-530. http://portal.acm.org/citation.cfm?id=972490

Holland, J.H. and J.H. Holland, 1975. Adaptation in Natural and Artificial Systems: An introductory analysis with applications to biology, control, and artificial intelligence. The University of Michigan Press, Ann Arbor, ISBN: 0472084607, pp: 183.

Hong, J.H. and S.B. Cho, 2006. Efficient huge-scale feature selection with speciated genetic algorithm. Patt. Recogn. Lett., 27: 143-150. DOI: 10.1016/J.PATREC.2005.07.009

John, G.H., R. Kohavi and K. Pfleger, 1994. Irrelevant Feature and the Subset Selection Problem. Proceeding of the 11th International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA., pp: 121-129. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.6141&rep=rep1&type=pdf

Jurafsky, D., 2004. Pragmatics and Computational Linguistics, In: Handbook of Pragmatics, Horn, L. and G. Ward (Eds.). Wiley-Blackwell, Malden, MA., pp: 578-604. ISBN: 10: 0631225471

Jurafsky, D., E. Shriberg, B. Fox and T. Curl, 1998. Lexical, prosodic and syntactic cues for dialog acts. Proceedings of ACL/Coling '98 Workshop on Discourse Relations and Discourse Markers, Montreal, Quebec, Canada, pp: 114-120. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.209

Kats, H., 2006. Classification of user utterances in question answering dialogues, Master's Thesis, University of Twente, Netherlands.

Keizer, S. and R. op den Akker, 2007. Dialogue act recognition under uncertainty using bayesian networks. Nat. Language Eng., 13: 287-316. DOI: 10.1017/S1351324905004067

Kelly, J.D. and L. Davis, 1991. Hybridising the Genetic Algorithm and the K Nearest Neighbors Classification Algorithm. In ICGA, pp: 377-383.

Kira, K. and L. Rendell, 1992. The feature selection problem: Traditional methods and a new algorithm. Proceedings of the National Conference on Artificial Intelligence, (NCAI'92), John Wiley and Sons Ltd., USA., pp: 1-29. http://direct.bl.uk/bld/PlaceOrder.do?UIN=000587606&ETOC=EN&from=searchengine

Kohavi, R. and G. John, 1997. Wrappers for feature subset selection. Art. Intell., 97: 273-324. DOI: 10.1016/S0004-3702(97)00043-X

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, ISBN: 0262111705, pp: 833.

Lal, T.N., O. Chapelle, J. Weston and A. Elisseeff, 2006. Embedded Methods. In: Feature Extraction: Foundations and Applications, Guyon, I., S. Gunn, M. Nikravesh and L.A. Zadeh (Eds.). Springer, Berlin, Germany, pp: 137-165. ISBN: 10: 3540354875

Langley, P., 1994. Selection of relevant features in machine learning. Proceedings of the AAAI Fall Symposium Relevance, (FSR'94), AAAI Press, USA., pp: 140-144.http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.9852

Lanzi, P., 1997. Fast feature selection with genetic algorithms: A filter approach. Proceedings of IEEE International Conference on Evolutionary Computation, Apr. 13-16, IEEE Xplore Press, Indianapolis, IN., USA., pp: 537-540. DOI: 10.1109/ICEC.1997.592369

Lecocke, M. and K. Hess, 2007. An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data. Can. Inf., 2: 313-327. PMCID: PMC2675488

Lee, S.H., H.I. Koo and N.I. Cho, 2010. Image segmentation algorithms based on the machine learning of features. Patt. Recogn. Lett, 31: 2325-2336. DOI: 10.1016/j.patrec.2010.07.004

Lesch, S., 2005. Classification of Multidimensional Dialogue Acts using Maximum Entropy DIploma Thesis. Saarl and University, Postfach 151150, D-66041 Saarbrucken, Germany.

Li, L., C.R. Weinberg, T.A. Darden and L.G. Pedersen, 2001. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. Bioinformatics, 17: 1131-1142. DOI: 10.1093/BIOINFORMATICS/17.12.1131

Li, Y., Z. F. Wu and J. M. Liu, 2004. Feature selection for high-dimensional data using two-level filter. Proceedings of the third International Conference on Machine Learning and Cybernetics, Aug. 26-29, IEEE Xplore Press, USA., pp: 1711-1716. DOI: 10.1109/ICMLC.2004.1382051

Liu, H. and H. Motoda, 1998. Feature Selection for Knowledge Discovery and Data Mining. Springer, New York, ISBN: 079238198X, pp: 214.

Liu, H. and L. Yu, 2005. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowledge Data Eng., 17: 491-502. DOI: 10.1109/TKDE.2005.66

Liu, W., M. Wang and Y. Zhong, 1995. Selecting features with genetic algorithm in handwritten digit recognition. Proceedings of the International IEEE Conference on Evolutionary Computation, Nov. 29-1 Dec., IEEE Xplore Press, Perth, WA, Australia, pp: 396-399. DOI: 10.1109/ICEC.1995.489180

Lu, J., T. Zhao and Y. Zhang, 2008. Feature selection based-on genetic algorithm for image annotation. Knowledge-Based Syst., 21: 887-891. DOI: 10.1016/J.KNOSYS.2008.03.051

Marquez, L., 2000. Machine Learning and Natural Language Processing. Technical Report LSI-00-45-R, Departament de Llenguatges i Sistemes Informatics, Universitat Politecnica de Catalunya.

Mitchell, M., 1996. An Introduction to Genetic Algorithms. MIT Press, Cambridge, ISBN: 0262133164, pp: 205.

Morariu, D.I., L.N. Vintan and V. Tresp, 2006. Evolutionary feature selection for text documents using the SVM. Proc. Worl Acad. Sci. Tech., 15: 215-221. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.100.8408&rep=rep1&type=pdf

Moser, A. and M.N. Murty, 2000. On the scalability of genetic algorithms to very large-scale feature selection. Real-World Appl. Evolut. Comput., 1803: 309-311. DOI: 10.1007/3-540-45561-2_8

Ozdemir, M., M.J. Embrechts, F. Arciniegas, C.M. Breneman and L. Lockwood *et al*., 2001. Feature selection for in-silico drug design using genetic algorithms and neural networks. IEEE Mountain Workshop on Soft Computing in Industrial Applications, June 25-27, IEEE Xplore Press, Blacksburg, VA, USA., pp: 53-57. 10.1109/SMCIA.2001.936728

Punch, W.F., E.D. Goodman, M. Pei, L. Chia-Shun and P. Hovl and *et al*., 1993. Further research on feature selection and classification using genetic algorithms. Proceedings of the 5th International Conference on Genetic Algorithms, (ICGA'93), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., 557-564. http://portal.acm.org/citation.cfm?id=657588

Saeys, Y., I. Inza and P. Larraaga, 2007. A review of feature selection techniques in bioinformatics. Bioinformatics, 23: 2507-2517. DOI: 10.1093/bioinformatics/btm344

Sajn, L. and M. Kukar, 2010. Image processing and machine learning for fully automated probabilistic evaluation of medical images. Comput. Methods Programs Biomed., DOI: 10.1016/J.CMPB.2010.06.021

Samuel, K., S. Carberry and K. Vijay-Shanker, 1999. Automatically selecting useful phrases for dialogue act tagging. Proceedings of 4th Conference of the Pacific Association for Computational Linguistics, (PACLING'99), Waterloo, Ontario, Canada, 1-14. http://adsabs.harvard.edu/abs/1999cs........6016S

Sanchez-Ferrero, G.V. and J.I. Arribas, 2007. A statistical-genetic algorithm to select the most significant features in mammograms. Proceedings of the 12th International Conference on Computer Analysis of Images and Patterns, (CAIP'07), Springer-Verlag Berlin, Heidelberg, pp: 189-196. http://portal.acm.org/citation.cfm?id=1770931

Schtz, M., 1997. Other Operators. Gene Duplication and Deletion. In: Handbook of Evolutionary Computation C3, Bck, T., Z. Fogel and D.B. Michalewicz, Hrsg (Eds.). Institute of Physics Publishing, New York, pp: 8-15. ISBN: 0750303921

Searle, J.R., 1975. A Taxonomy of Illocutionary Acts. In: Language, Mind and Knowledge, Minnesota Studies in the Philosophy of Science, Gunderson, K. (Eds.). U of Minnesota Press, Minnesota, pp: 344-369. ISBN: 0816657793

Shahla, N., M.E. Basiri, N. Ghasem-Aghaee and M.H. Aghdam, 2009. A novel ACOGA hybrid algorithm for feature selection in protein function prediction. Expert Syst. Appl., 36:12086-12094 DOI: 10.1016/J.ESWA.2009.04.023

Siedlecki, W. and J. Sklansky, 1988. On automatic feature selection. Int. J. Patt. Recog. Art. Intell., 2: 197-220. DOI: 10.1142/S0218001488000145

Silla, C.N., G.L. Pappa, A.A. Freitas and C.A.A. Kaestner, 2004. Automatic text summarization with genetic algorithm-based attribute selection. Adv. Art. Intell. IBERAMIA, 3315: 305-314. DOI: 10.1007/978-3-540-30498-2_31

Smith, S.F., 1980. A learning system based on genetic adaptive algorithms. Doctoral Dissertation Thesis. University of Pittsburgh Pittsburgh, PA, USA. http://portal.acm.org/citation.cfm?id=909835

Sridhar, V.K.R., S. Bangalore and S. Narayanan, 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. Comput. Speech Language, 23: 407422. DOI: 10.1016/j.csl.2008.12.001

Vafaie, H. and K. De Jong, 1992. Genetic algorithms as a tool for feature selection in machine learning. Proceedings of the 4th International Conference on Tools with Artificial Intelligence, Nov. 10-13, Arlington, VA. USA., pp: 200-203. DOI: 10.1109/TAI.1992.246402

Vafaie, H. and K. De Jong, 1995. Genetic algorithms as a tool for restructuring feature space representations. Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, Nov. 5-8, Henidon, VA. USA., pp: 8-11. DOI: 10.1109/TAI.1995.479372

Verbree, A.T., R.J. Rienks and D.K.J. Heylen, 2006. Dialogue act tagging using smart feature selection: Results on multiple corpora. Proceedings of the International IEEE Spoken Language Technology Workshop, Dec. 10-13, IEEE Xplore Press, Palm Beach, pp: 10-13. DOI: 10.1109/SLT.2006.326819

Vose, M.D., 1999. The Simple Genetic Algorithm: Foundation and Theory, MIT, Press, Cambridge, Mass. London, ISBN: 026222058X, pp: 251.

Webb, N., M. Hepple and Y. Wilks, 2005. Dialogue act classification based on Intra-utterance features. University of Sheffield, UK. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.80.8356&rep=rep1&type=pdf

William, H., 2004. Genetic wrappers for feature selection in decision tree induction and variable ordering in Bayesian network structure learning, Inform. Sci. Int. J., 163: 103-122. 10.1016/j.ins.2003.03.019

Xing, E.P., M.I. Jordan and R.M. Karp, 2001. Feature selection for high-dimensional genomic microarray data. Proceeding of the 8th International Conference on Machine Learning, (ICML'01), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., pp: 601-608.

Yahya, A. A, R. Mahmod, F. Ahmad and N. Sullaiman, 2006. Dynamic Bayesian networks for intention recognition in conversational agent. Proceedings of the 3rd International Conference on Artificial Intelligence in Engineering and Technology (iCAiET2006), University Malaysia Sabah, Sabah, Malaysia.

Yahya, A.A., A.R. Ramli and T. Perumal, 2009. Dynamic bayesian networks with ranking-based feature selection for dialogue act recognition. Proceedings of the ATCi Advanced Technology Congress 2009, Nov. 3rd -5th, PWTC, Kuala Lumpur.

Yang, A., D. Li and L. Zhu, 2011. An improved genetic algorithm for optimal feature subset selection from multi-character feature set. Expert Syst. Appl. Int. J., 38: 2733-2740. DOI: 10.1016/j.eswa.2010.08.063

Yang, J. and V. Honavar, 1998. Feature subset selection using a genetic algorithm. IEEE Intell. Syst., 13: 44-49. DOI: 10.1109/5254.671091

Yu, E. and S. Cho, 2003. GA-SVM wrapper approach for feature subset selection in keystroke dynamics identity verification. Proceedings of the IEEE International Joint Conference on Neural Networks, July 20-24, IEEE Xplore Press, pp: 2253-2257. DOI: 10.1109/IJCNN.2003.1223761

Yu, L. and H. Liu, 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. Proceedings of the International Conference on Machine Learning (ICML-03), AAAI Press, Washington, DC., pp: 856-863. http://www.aaai.org/Papers/ICML/2003/ICML03-111.pdf

Zebulum, R.S., M. Vellasco and M.A. Pacheco, 2000. Variable length representation in evolutionary electronics. Evol. Comput. J. 8: 93-120. DOI: 10.1162/106365600568112

Zhang, C.K. and H. Hu, 2005. An effective feature selection scheme via genetic algorithm using mutual information. Fuzzy Syst. Knowl. Discovery, 3614: 73-80. DOI: 10.1007/11540007_10

Zhang, L., J. Wang, Y. Zhao and Z. Yang, 2003. A novel hybrid feature selection algorithm: Using ReliefF estimation for GA-Wrapper search. Proceedings of IEEE International Conference on Machine Learning and Cybernetics. Nov. 2-5, IEEE Xplore Press, USA., pp: 380-384. DOI: 10.1109/ICMLC.2003.1264506

Zhang, P., B. Verma and K. Kumar, 2004. A neural-genetic algorithm for feature selection and breast abnormality classification in digital mammography. IEEE, 3: 2303-2308. DOI: 10.1109/IJCNN.2004.1380985

Zheng, H. and Y. Zhang, 2008. Feature selection for high dimensional data in astronomy. Adv. Space Res., 41: 1960-1964. DOI: 10.1016/J.ASR.2007.08.033

Zhuo, L., J. Zheng, F. Wang, X. Li and B. Ai *et al*., 2008. A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. Remote Sens. Spatial Inform. Sci., 7147: 397-402. http://www.isprs.org/proceedings/XXXVII/congress/7_pdf/3_WG-VII-3/32.pdf

Zhu, H., L. Jiao and J. Pan, 2006. Multi-population genetic algorithm for feature selection. Adv. Nat. Comput., 4222: 480-487. DOI: 10.1007/11881223_59