Optimal Feature Selection for Data Classification and Clustering: Techniques and Guidelines

Farhad Rad
yasooj branch, islamicazad university

Ali AsgharNadri
yasooj branch, islamicazad university

Hamid Parvin
yasooj branch, islamicazad university

ABSTRACT
In this paper, principles and existing feature selection methods for classifying and clustering data be introduced. To that end, categorizing frameworks for finding selected subsets, namely, search-based and non-search based procedures as well as evaluation criteria and data mining tasks are discussed.In the following, a platform is developed as an intermediate step toward developing an intelligent feature selection system, involving crucial, decisive and effective factors in feature selection process. The procedure increases accuracy in classification and goodness of clusters. Finally, some of the problems and challenges facing the current and future feature selection processing are also discussed.

Keywords
Feature selection, classification, clustering,categorizing framework, evaluation criteria.

1. Introduction
Despite recent advances in data processing technology, the ever-increasing volume of datasets and the number of features, which often increase waste of data, makes it difficult to supply the needed resources, including storage and processing of data. At the same time, the explosive growth of data has led to increased noise. The misleading and waste data among the mass of useful data, exchanged in social networks, are the manifestations of noise in the data. Curse of dimensionality is the most important outcome of increasing data dimensions. Additionally, despite the large number of features, learning models are prone to overfitting and performance degradation [1]. Several strategies have been proposed in the literature to deal with such consequences. For example, qualitative and targeted data reduction can help to solve the problem in a more limited scale, without removal of useful or meaningful data. Some strategies proposed for data reduction are dimension reduction, data reduction, and data compression[2]. In this article, the authors focus on dimension reduction techniques, which are among common techniques for reducing the number of features,.

1.1 Feature Extraction
In this method, some prominent features are produced through one or more conversions on input features. While mapping points from one space with higher dimensions into another space with lower dimensions, a large number of points may overlap. Feature extraction helps to find a new dimension where a minimum number of points may overlap. This approach is associated with the problem area and is commonly used in image processing where specific features are extracted in accordance with the requirements of the problem.

1.2 Feature Selection
Proposed in various fields of machine learning and data mining, feature selection is one of the subsidiaries of feature extraction. It is preferable in contexts where readability and interpretability are issues of concern, because the discounted values of the main features are preserved in the reduced space [1]. This method of dimension reduction results in a qualitative database, without removal of useful information. It also allows for the features with different data models to be combined. The issue is of importance because a large number of features are often used in different applications. Therefore, the need to select a limited set from among them becomes apparent. Constraints and considerations such as avoiding the curse of dimensionality, memory limits, reduction of the needed computations, and reduction of the runtime, among others oblige one to select the minimum number of features to be used in prediction of future data.
The following general objectives and considerations should be taken into account considering feature selection as well as and the characteristics that the final subset should have:

1. Idealization: a minimally sized feature subset that is necessary and sufficient to the target concept should be found [3].
2. Optimalityof the final subset: the amount obtained by

evaluation function for a given subset should be maximal compared to other subsets.

3. Approximating original class distribution: the goal of feature selection is to select a small subset such that the resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution given all feature values[3].

4. Increased distinctiveness: for clustering, the most effective subsets (of the main features), which lead to increased distinctiveness of the clusters are selected, because some features are not effective for cluster building.

5. Dealing with bias: the selected features should represent the hidden knowledge into the problem and highlight the hidden biases in the features and datasets.

6. Maintaining accuracy: the classification accuracy should not significantly decrease. Reduction in the size of the structure should be achieved without any significant reduction in the accuracy prediction.

A caveat is in order here: there is not a guarantee that the parameters which are optimized for the full set of features are equally optimal for the final subset of the features. Feature selection has its own advantages and disadvantages to be briefly reviewed bellow:

• Advantages are[4]:

1) Adjustment of curse of dimensionality issue and avoidance of overfitting and improved performance of the model,

2) Dimension reduction of feature space, reduction of computations and memory requirements,

3) Improvement in visualization and data understanding as well as better data comprehensibility,

4) Reduction of training time and final model utilizing time and increase of speedup of data mining algorithm.

5) Removal of redundant and irrelevant features or noisy data and increase of accuracy of the final model with a focus on potentially useful features in order to improve the quality of data,

6) Reduction of the number of features, which leads to reduced need for storage and efficient use of resources in the next rounds of data processing.

7) Comprehension of data behavior and detection of hidden patterns in data as well as obtaining a deeper insight of the data generation processes that lead to the acquisition of knowledge, and Provision of more cost effective models.

• Disadvantages are:

1) Attempting to find an effective and representative subset of the available features as an additional step in data mining,

2) Searching the existing virtual space, which adds another dimension to the issue, namely, finding a subset of optimal relevant features, and Spending additional time in the learning phase.

4) This step is an exploratory process (delving into various aspects of the problem to isolate the most important dimensions and the most effective).

Feature selection process is divided into four parts: generation procedure, evaluation function, stopping criteria and validation procedure.

The rest of the article is organized as follows:

In section 2, Selecting a subsetis reviewed briefly. In section 3, Methods of feature selection are described. The proposed developed platform is presented in section 4.Existing and future challenges to feature selection in the field of research and development are summarized in section 5 and finally conclusion is summarized in section 6.

2. Selecting a Subset (Generation Function)

2.1 Search-based strategies

This is also referred to as subset generation in which a candidate subset is determined for evaluation in each state of the search space. Two most important issues at this stage are as follows:

a) Successor generation at every step: This step has to do with deciding about the starting point of the search that determines its trajectory [5]. That is to say, in every situation, forward, backward, compound, weighting and the random strategies can be used to decide on possible starting points of search as explicated below[5]:1) Forward strategy: in this strategy, the selected subset is first empty and later the features that are not selected yet are added to the selected subset in case they can decrease the classification error rate. The process continues in the same manner until the final subset is selected. 2) Backward strategy: unlike the forward strategy, in this strategy first all features are considered as the selected subset. Then the features that decrease the classifier error rate are removed from the selected subset. The process continues until the final subset is obtained. 3) Compound strategy: this strategy is to do with k sequential forward steps and L sequential backward steps. That is to say, forward or backward steps are performed in order to discover new interactions occurring between features. 4) Random strategy: it generates a random mode at any stage. Other operators are limited with some criteria such as the number of features or error rate (in each stage).5) Weighing strategy: It assigns weights to all features of the desired solution iteratively until all possibilities have been taken care of based on repetitive sampling of the existing set of samples.

b) The supervision of the feature selection process: The Search organization supervises the feature selection process through heuristic, complete, and random strategies. The following are the main types of this process.

1) Exponential search: it is an optimized search-based strategy through which we can obtain the optimal subset.

2) Sequential (heuristic) search: This procedure allows for a feature to be selected from among all the next features repeatedly. The number of possible steps in this phase is O(N). Easy implementation is the main advantage of this method. Sequential forward search (SFS), sequential backward search (SBS) and, the bidirectional selection are examples of this method [6].

3) Non-sequential (random) search: This method starts with a randomly selected subset to achieve the optimized subset either through successor completely random subset generation as in Las Vegas or through a kind of sequential search that applies randomness (avoidance of local optima in the search space) in the sequential search approach [6].

4) Genetic algorithm: It is a flexible and powerful technique of random search for finding approximate solutions to optimization and search problems. It uses pattern matching to find the optimal solution and genetic evolution as ways to solve problems. The problems dealt with by this algorithm have to do with inputs turning into a solution through a process modeled on genetic evolution. An evaluation function built in this algorithm evaluates the solutions as candidates and in case the exit condition is met, the evaluation process comes to an end. The algorithm allows for the most appropriate solutions rather than the best ones to be selected. There are some uncertainties about the algorithm, including whether the algorithm would become convergent or it wouldmove towards a good solution. Local optima are also another uncertainty in this algorithm.

2.2. Non-search Strategies:

These strategies, also referred to as mathematical optimization, aim at reducing redundancy of the features and maximizing the relationship between the feature and the target variable. A mathematical search uses the necessary and sufficient conditions that are proved to be true for the answer for the optimization. The overall goal of the optimization is to find the best acceptable solution while taking the constraints and needs of the problem into account. Therefore, the selection of an appropriate objective function is one of the most important steps in optimization. There are two types of optimization:

1) Combinatorial Optimization method: a branch of optimization which deals with the optimization problems that are generally difficult to solve. These issues are usually solved in the best possible way by efficient examination of a (usually large) space of all possible answers. Travelling Salesman Problem (TSP), Quadratic Assignment Problem (QAP), timeline and scheduling issues are examples of Combinatorial Optimizationissues.

2) Game theory method: a technique for finding the selected subset that solves the problem through different methods more efficiently. It also delivers better results. This approach is known as the science of interactive decision-making. In this case, the problem is first simulated in a solvable manner through the concepts of game theory. Problem environment is presented in form of a game that can be navigated in different ways. Finally, the optimal or semi-optimal subsets of features are selected. Learning from the environment takes place while navigating the path, allowing for the best regions for navigation to be selected in the future repetitions [1]. Reinforcement learning and Monte Carlo methods are used to navigate the environment in this way. Each subset of the total features is a state of state space and adding each feature to the set is also considered as an act that leads us towards a new state. In fact, each state is an unseen state relative to the past. At each stage, one feature that has obtained the highest points in the previous steps and has led the issue towards a better state is selected.

3. Methods of Feature Selection

Filter Method: As an ad-hoc preprocessing feature selection method, it takes into account the general characteristics of the data such as information gain or statistical dependencies regardless of the evaluation criteria to select the subset of features. An Independent criterion is used to evaluate the subset of features. This method is computationally fast and is dependent on the used classifier. In some cases, it eliminates useless features, while these features may be useful in combination with other features. Since microarray data analysis is one of the most common uses of the feature selection, this method is used to analyze the microarray data. Feature selection for this type of data is done in two ways as follows [6]:

Ranking Methods: Most filter methods view the issue of feature selection as a ranking issue. Univariate and bivariate filter methods are ranking methods most often used for micro-array data analysis:

1) Univariate methods: They are subdivided into two groups, namely, parametric and non-parametric methods in accordance to their objectives. It is also noteworthy that in some cases due to the intrinsic importance of the features and regardless of their likely dependence to one another the univariate filter methods are used.

• *Parametric methods:* they allow for the data to be drawn from a given probability distribution, on these some more or less explicit assumption.

• *Non-parametric methods:* they allow for data to be drawn on the bases of some unknown distribution. To quantify the difference in expression between classes based on some estimate scoring function is used.

2) Bivariate methods: in accordance with their discrimination power between two or more conditions, ranking pairs of genes can be performed either using a “greedy

strategy" or "all pair strategy."

• *Greedy strategies:* they first rank all genes by individual ranking, using one of the criteria provide by univariate ranking methods; subsequently, the highest scoring gene is paired with the gene gj that gives the highest gene pair score. After selecting first pair, gs that is next highest ranked gene paired with the gene gr, it maximizes the pair score, and so on.

• *All-pairs strategy:* unlike the greedy methods, all pair's strategies compute the pair scores for all pairs. As such, they examine all possible gene pairs.

- Filter Methods-Space Searching Approach

It is an optimization strategy that will provide the most informative and least redundant subset of features among the whole set. This strategy follows three main steps described below:

●Multivariate methods. In this method, the features are evaluated batches and in comparison to other features, that's why this method is able to identify and control redundant features.

- Wrapper Method: the model uses the prediction accuracy of the algorithm used to determine the quality of the selected features and acts as a black box; it means that it has no parameter exchange with the outside world, and represents the simplicity of feature selection procedure in this approach. The Algorithm also serves as evaluation criterion for features subset. The main disadvantage of this method is its high complexity that leads to uncontrollability of problem solving. In other words, the mining algorithm controls the selection of features subset. Feature selection is done with immediate intervention of the classifier and based on its evaluation and the features subsets are introduced in accordance with their usefulness for the predictor. The search for finding a good feature subset is done by the algorithm itself as a part of the evaluation function. This method results in performance than the filter methods, but training sets with low volume usually tend to result in overfitting.

Wrapper methods for feature selection are either *Deterministic* or *randomness.* In the deterministic method (examples: SFS-SBE-Beam Search), the search starts from the existing feature space, in forward or backward manners [7]. Randomness (examples SA-GA- Random Hill Climbing): Compared to the deterministic method, the next features subset is randomly searched.

- Hybrid Method: this method is used when you want to have an additional evaluation of the features, or attend the samples behavior more effectively. Although, the calculations of this method are more than the calculations in the filter and wrapper method, this method delivers more accurate results for classification and goodness cluster for clustering. This method is generally used for large datasets and more complex goals. Combinatorial algorithms are proposed to take advantage of the above-mentioned models and to avoid pre-description of a stopping criterion, in large datasets [4]. The accuracy of this method is usually comparable to the accuracy of the wrapper method and its performance is comparable to that of the filter method. The quality of the clustering algorithm results in this method will determine the stopping condition.

- Embedded Method: a feature of this method is the adjustment of the feature selection process. Feature selection is carried out implicitly within the classifier context and is considered as a part of the training procedure. While classifying, the algorithm (itself) decides which features to use or ignore. Thus, it is difficult to control the selection of appropriate number features, often resulting in redundancy of features. The predictive model is generated and trained along with the feature selection. This approach would be more effective if the user has an initial insight of the subject and its intended use.

4. Proposed Developed Platform

This study proposes, generalizes, and designs an original platform to develop the mentioned categorizing method by introducing more dimensions from the perspective of the user [48]. The platform proposes that effective factors in feature selection should be divided into two groups: ***knowledge*** and ***data***. The items related to each of them are separately explained in follow.

The ***knowledge*** factor includes elements such as the ***purpose*** of feature selection, the ***time*** required to achieve it, ***logic*** of problem, the type of ***expected output,*** the ***rate of M/N*** (***M***=the number of selected features; ***N***=the total number of features), the I/O ***stream***, and ***relevance*** of the features. The data factor includes elements such as ***class information, the type of features,*** the quality of the used data, the number of samples (the ratio of the number of features to the number of samples) and the circumstances of describing data clusters. Our overall goal is to optimize the process, including the procedure of its implementation, the quality of results and accuracy of the features.

In the following, the role of these factors in selection of the appropriate algorithm or design new algorithms is explained. The proposed platform be shown in figure1 and highlights two goals: (1) introducing the existing algorithms with similar characteristics and checking their strengths and weaknesses,(2) providing guidelines for designing an intelligent feature selection system.

After the presentation of the concepts and definitions of the issue, it is understood that the basic arrangements for designing an intelligent system for intelligent feature selection are automatically provided. The role of the domain expert, users, or availability of knowledge domain on each of the subsections of these two factors can lead to reduction of the feature selection process time as well as the time for selection or design of a suitable algorithm

for the intended purpose. In [5]the author explains that the new feature selection methods increasing to tackle specific problems with different strategies. The general objectives of the process are (1) ensemble method ensure a better behavior of feature selection, (2) using this method in combination with other techniques such as tree ensemble and feature extraction, (3) to reinterpret the existing algorithms, (4) to create a new method to deal with still-unresolved problems, and (5) to combine several feature selection methods.

5.Existing and future challenges to feature selection in the field of research and development

Scalability of Algorithm: in the data preprocessing phase (feature selection), the classical algorithms should be changed in such a way that they can scan multiple databases and or result in improved access to the data. However, along with the increased size of the datasets, the issue of scalability (in the existing algorithms) becomes more problematic. Some existing methods of feature selection for classification should also save all the dimensions in the memory. Usually, they require a sufficient number of samples to obtain, statically, adequate results. Some methods also try to overcome this problem by allocating memory (only to minor samples or those without adequate information). In conclusion, we believe that the scalability of classification and feature selection methods should be given more attention to keep pace with the growth and fast streaming of the data [3].

• ***Robustness of Algorithm:*** if the algorithm is not robust, with any change in training samples per execute, users receive different results from its implementation. Another problem occurring in the machine learning is when in face of volatile data, the results of the algorithm, based on the number of samples of various classes, are equal to each other. Here modeling should be carried out consistently with increased testing of the produced model. It is assumed that the model is less robust.

• ***Stability:*** it refers to sensitivity of selection in the face of data perturbation that is in the context of future data samples. Stability leads to reliability of the results obtained from that. However, stability (alone) will not be sufficient. This measure represents the insensitivity of the algorithm to different features of the training set, and depends not only on the learning algorithms, but also on the dataset. Stability of a feature selection algorithm can be viewed as the consistency of an algorithm to produce a consistent feature subset when new training samples are added or when some training samples are removed [8].

• ***Small Sample Size:*** in some applications, it is observed that the number of samples in the dataset is very small. Given that the samples indicate the knowledge existing in the problem, in their absence, machine's learning process will be done very slowly or the final model is not reliable and pervasive.

• ***Selection of Appropriate Evaluation Criteria:*** in some applications, the selection of appropriate evaluation criteria is challenging. This is affected by the dataset is available. In some cases, the use of another evaluation method (for a particular data mining method), gives greater desirability to the target. For example, for a special task, the use of the correlation measure instead of the distance measure in filter algorithms may impose fewer calculations.

• ***Approaches to Learning:*** It refers to the degree of supervision or structure imposed on the learning process. There are five approaches to learning as follows: (1) supervised learning, otherwise known as classification, (2) unsupervised learning which is used to identify data groups where no clear image of the identified clustered may be achieved before the label of the data is determined, (3) semi-supervised learning where the labeled samples are used to train the classes model and the unlabeled samples are used to identify the bounds of clusters, (4) active learning where the users or experts themselves label the specific samples, and (5) transfer learning, in which knowledge is derived from one or more reference model that is used to create an objective model [2]. Obviously, users have a difficult time deciding which approach to learning to choose.

• ***Analysis of Social Networks:*** The analysis of online communities such as Facebook has also become increasingly important as one of the data mining applications. Due to the dispersion of knowledge in problem, extraction of useful knowledge in these areas is very difficult.

• ***Feature Selection in Ultrahigh-Dimensional Contexts:*** in some cases, different approaches can be used to select effective qualitative dimensions that provide valuable concepts when put together, thereby reducing the number of dimensions. Due to the close relationship among the dimensions of the interest, in some applications, the removal of one or more dimensions may cause a break in the acquisition of useful knowledge from datasets. Thus, intermediate knowledge of these dimensions may be ignored. Ultrahigh dimensionality not only incurs unbearable memory requirements and high computational cost in training phase but also worsens their generalization ability because of the phenomenon known as curse of dimensionality [9].

• ***Feature Selection in Active Samples:*** different sampling methods lead to adjustment of knowledge in the datasets and reduce the robustness of the model. Typically, random sampling is used in very large training datasets. However,Feature Selection in Active Samples avoids pure random sampling. Instead, selection is realized by selective sampling that takes advantage of data characteristics when selecting instances [4]. Thus, an attempt is made

to select that portion of the samples which are more informative, or at the center of attention in evaluation of relevance of features.

• ***Subspace Search and Sample Selection:*** in clustering, many clusters may exist in different subspaces for small dimensionality with overlapped or non-overlapped dimensions [5]. Therefore, subspace searching has not only to do with the feature selection problem but also it has to do with finding many subspaces in which feature selection finds one subspace. Therefore, it provides an efficient algorithm for clustering.

• ***Feature Selection with Sparse Data Matrix:*** sparse data refers to a relatively high percentage of variables that do not have actual data. There are two types of sparsity, namely, controlled sparsity and random sparsity [5]. Controlled sparsity refers to a range of values of one or more than one dimensions that have no data [84]. Random sparsity, in contrast, refers to empty values scattered throughout the data variable.

• ***Linked Data:*** theyare among other challenges of feature selection for classification. Feature selection methods for linked data need to solve the following immediate challenges: (i) how to exploit relations among data instances; and (ii) how to take advantage of these relations for feature selection [1]. In linked data, the data structure is linked in such a degree that it is difficult to analyze for data mining.

• ***Feature selection in new applications:*** it is observed that a very large volume of data is stored in databases, while tools for processing and exploiting them, and pre-processing methods in particular are required. Storing data also increases computational costs and learning time. Using other methods for data reduction also leads to ignoring the useful data or makes it more difficult to identify favorable data for intended uses.

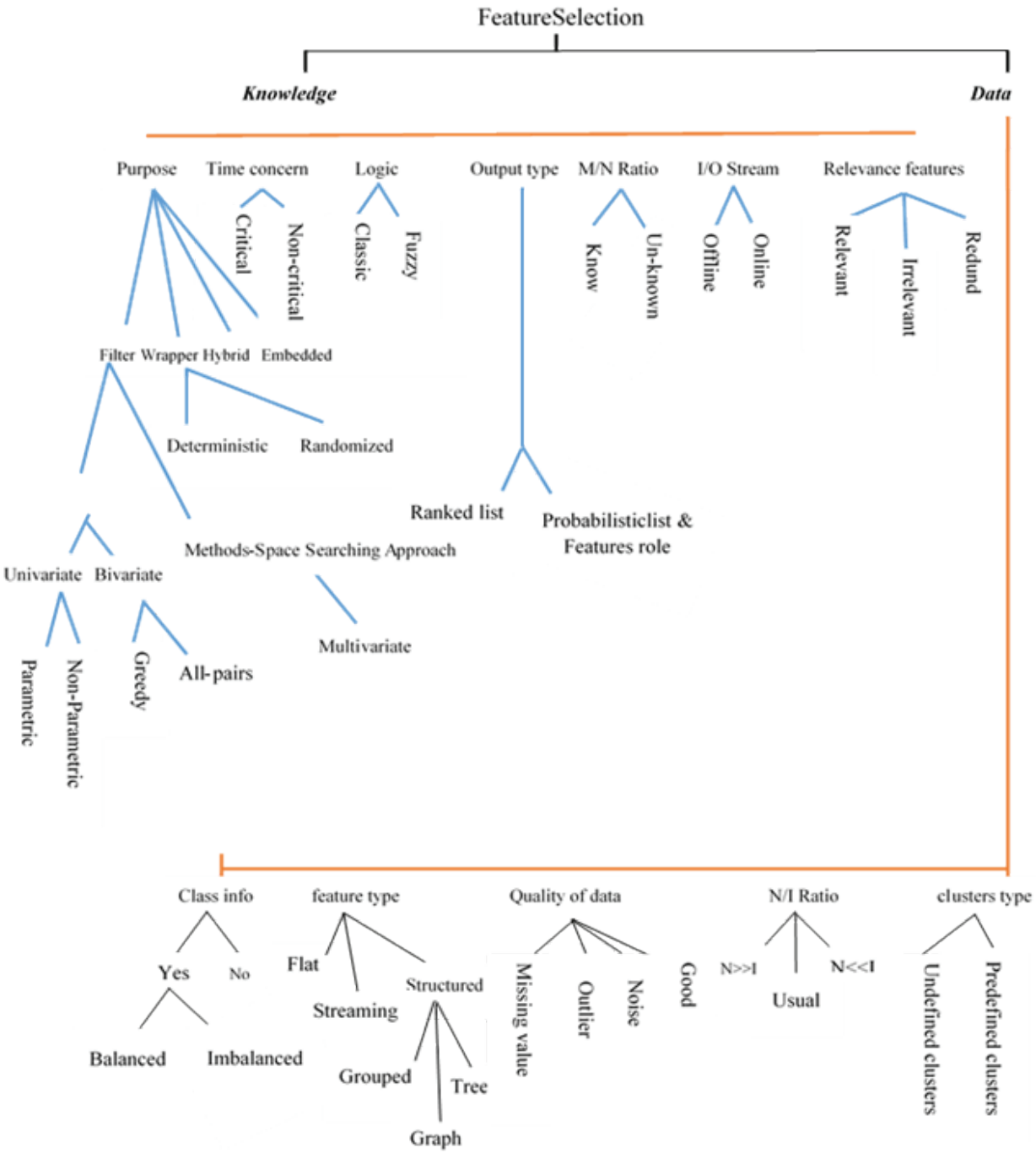
• ***Data Reduction in the Entry of Data StreamsPhase:*** new feature selection applications should include dynamic methods for this purpose, especially in a situation in which our ability to store data surpasses our processing and exploitation capabilities. This process aims to pay more attention relevant datasets to improve data processing.

• ***Feature Selection in Distributed Systems:*** systems such as Grid and Cloud are common examples of distributed systems, where data resources or data files are distributed. In some cases, the data should be collected from heterogeneous environments, and must be adapted to the target platform, exemplifying risks in the nature of feature selection process. Sometimes the data in different sectors are significantly different in terms of structure and quality, and their preprocessing is time-consuming. Such facts may prevent consistent feature selection.

6. Conclusion

In this paper, a structure was defined in order to design an intelligent feature selection system which provides a suitable feature selection algorithm for each application and in accordance with the effective factors, or if needed, can design a suitable feature selection algorithm for the intended application. In this regard, at first a three-dimensional categorization consisting of strategies for finding selected subsets, evaluation criteria and data mining tasks was developed. Each categorizing block contains algorithms with similar characteristics. Then a unifying platform was developed, which is the foundation of designing an intelligent system for selecting an appropriate algorithm for each application with each of these factors. Finally, challenges in the research and development of feature selection were introduced so that the problems and concerns related to this topic may be taken into consideration in further studies.

Figure 1: Proposed Developed Platform



Reference

1) S. Alelyani, J. Tang, and H. Liu, "Feature Selection for Clustering: A Review," *Data Clustering: Algorithms and Applications*, vol. 29, 2013.

2) J. Han and M. Kamber, "Data mining: concepts and techniques (the Morgan Kaufmann Series in data management systems)," 2000.

3) M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, pp. 131-156, 1997.

4) H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 17, pp. 491-502, 2005.

5) V. Kumar and S. Minz, "Feature Selection," *Smart* CR, vol. 4, pp. 211-229, 2014.

6) V. Tyagi and A. Mishra, "A Survey on Different Feature Selection Methods for Microarray Data Analysis," *International Journal of Computer Applications,* vol. 67, pp. 36-40, 2013.

7) L. Ladha and T. Deepa, "Feature selection methods and algorithms," *International journal on computer science and engineering*, vol. 3, pp. 1787-1797, 2011.

8) G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* vol. 40, pp. 16-28, 2014.

9) M. Tan, I. W. Tsang, and L. Wang, "Towards ultrahigh dimensional feature selection for big data," *The Journal of Machine Learning Research,* vol. 15, pp. 1371-1429, 2014.