*I.J. Information Technology and Computer Science,* **2014, 11, 66-77**
Published Online October 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2014.11.10

# Feature Selection: A Practitioner View

Mr. **Saptarsi Goswami**
Assistant Professor, Institute of Engineering & Management, Kolkata, India
Email: Saptarsi.Goswami@iemcal.com

Dr. **Amlan Chakrabarti**
Associate Professor, A.K.Choudhury School of Information and Technology, Calcutta University, Kolkata, India
Email: acakcs@caluniv.ac.in

***Abstract—*** Feature selection is one of the most important preprocessing steps in data mining and knowledge Engineering. In this short review paper, apart from a brief taxonomy of current feature selection methods, we review feature selection methods that are being used in practice. Subsequently we produce a near comprehensive list of problems that have been solved using feature selection across technical and commercial domain. This can serve as a valuable tool to practitioners across industry and academia. We also present empirical results of filter based methods on various datasets. The empirical study covers task of classification, regression, text classification and clustering respectively. We also compare filter based ranking methods using rank correlation.

***Index Terms—*** Feature Selection, Supervised, Unsupervised, Commercial, Application Domain

## I. INTRODUCTION

There has been a surge in the dimension and the volume of datasets with information coming through mobiles, cameras, wireless sensory networks, radio frequency identification readers, weblogs, clickstreams, social media, internet search, video surveillance to name a few. The size of digital universe should have passed 1.8 zettabyte in 2011 as noted in [1]. World's per capita capacity to store digital information has roughly doubled every forty months [2]. Apart from the volume, the variety has added a lot of complexity to the analysis problem. Beside the structured data stored in traditional databases, we have unstructured data in form of text, video, images and semi-structured data in terms of XML and RSS feeds. Actually 90% of the data are unstructured as per [1]. For a holistic view, data mining and knowledge discovery on the unstructured data also needs to be carried out. This is bringing many new technologies like MapReduce, Columnar data store, Cloud at forefront. So we argue, dimensionality reduction and feature selection for data mining tasks have become more important in this changed 'input dataset scenario'.

Both feature selection and dimensionality reduction is effective in reducing the number of inputs without There has been a surge in the dimension and the volume of datasets with information coming through mobiles, cameras, wireless sensory networks, radio frequency identification readers, weblogs, clickstreams, social media, internet search, video surveillance to name a few. The size of digital universe should have passed 1.8 zettabyte in 2011[1]. World's per capita capacity to store digital information has roughly doubled every forty months [2]. Apart from the volume, the variety has added a lot of complexity to the analysis problem. Beside the structured data stored in traditional databases, we have unstructured data in form of text, video, images and semi-structured data in terms of XML and RSS feeds. Actually 90% of the data are unstructured [1]. So the potential of the unstructured data is immense. This needs to be tapped to have better models. Mostly the unstructured data also needs to bring into some sort of structure before inclusion in the models. One principal difference this 'structured data' will have much more dimensions than the usual structured data. As example a text document, is often represented as a Bag of Words (BoW). A document will be presented as a vector with thousand of features. So we argue, dimensionality reduction and feature selection for data mining tasks have become more important in this changed 'input dataset scenario'.

Both feature selection and dimensionality reduction is effective in reducing the number of inputs without compromising much on the information content and hence achieving mainly below three advantages [3], [4], [5]

(i) *Better model understandability and visualization* - It might not be possible to reduce to a two dimensional or a three dimensional feature set, but even if we want to visualize with combination of two or three features, the combinations will be much lesser.

(ii) *Generalization of the model* and reduced over fitting, as a result better learning accuracy is achieved.

(iii) *Efficiency in terms of time and space complexity* for both training and execution time.

In techniques like PCA (Principal Component Analysis) or factor analysis etc. we construct new features from the original features. Generally the no. of original features (n)and the no. of transformed features(n) are same. Rather than the entire set of transformed features we can select a few of them (k, $k < n$)which explains almost all of the variance of the original dataset. These families of methods are called *dimension reduction methods*.

 

Selecting few features from the original set of features based on measures like correlation, entropy and mutual information etc. Among the features and the target variable (if one exists) is called *feature selection*. Feature selection is also known as variable selection, attribute selection or variable subset selection. In case of dimensionality reduction, because of the transformed space, the newly constructed features are no more meaningful to the business community [6], or provide no opportunity to incorporate domain knowledge [4]. Also, it still does not eliminate the need of a collection of all the input attributes. For the above reasons, we stick to reviewing only feature selection.

In this paper, our specific contributions are as follows:

- We bring a practitioner's view, by comparing research state of the art and Industry state of the art
- We are comparing the feature selection metrics in terms of their similarity with each other, based on our empirical evaluation on various datasets.
- A thorough listing of problem areas across application domains and technical domains where feature selection has been used.

The organization of the paper is as follows: Section II provides a brief overview of the taxonomy of the feature selection methods and existing state of art. In Section III, we discuss about feature selection methods as available in commercial data mining packages. In Section IV, various available feature selection scoring methods are discussed. We summarize the problems solved by feature selection in various applications as well as technical domains in Section V. Section VI lists the results of our own empirical evaluation. This section consists of four tracks – text classification, classification, regression and clustering respectively. In section IV and section VI we have focused mainly on feature selection methods which rank the features, while in section III and section V we have given an overall view.

## II. Feature Selection

In this section, we briefly review the feature selection algorithms. Various dimensions of comparison are going to be: - output of the algorithm (subset or ranking), Supervised or Unsupervised (In terms of the task involved), input data (Continuous, discrete, binary, ordinary), Principle (Filter, Wrapper, Embedded). A search in Google Scholar for only year 2012 brings a result of 67,500 and a search in Arnetminer brings about 9715 publications. Above numbers surely reflect the amount of research done in this area. It's not possible to summarize all these findings in this review; we have looked at important papers in terms of their citations and focused mainly on last 10 years. Rather than a formal review, we try to find answers to the below questions:

- What is the research trend in feature selection?
- What is the state of the art in Industry in terms of feature selection? How is industry catching up with the cutting edge research?
- What are the various application domains and technical domains for feature selection?
- Which feature selection, scoring techniques (ranking) is similar?

The core belief or assumption of feature selection is that, not all features of an observation are equally relevant. There will be irrelevant features (with no information content relevant to the task) and redundant features (information represented by this feature, is already captured by other features). Removing them will potentially give a better generalization with less testing and training time and also better understanding and visualization. The problem of feature selection can be approached in various ways, which is described subsequently.

### *A. Different Approaches:*

As noted in [4],[5],[6] there are four standard approaches, which are briefed below:

*Embedded Approach*: In this approach, feature selection is a part of the objective function of the algorithm itself. Examples of the same are decision tree, LASSO, LARS, 1-norm support vector etc.

*Wrapper Approach:* In this method, the feature selection is approached considering the data mining algorithm as a black box. All possible combinations of the feature sets are used and tested exhaustively for the target data mining algorithm and it typically uses a measure like classification accuracy to select the best feature set. An obvious criticism is the time complexity of the method because of its 'brute force' nature. However, many heuristic and greedy methods can be used to prune the search space. To reemphasis, the optimal feature set will vary between algorithms. The techniques differ, in terms of strategy to search the feature space. The different techniques of searching the feature space are – Complete Search, Sequential Search and Randomized Search [3] respectively.

*Filter Approach:* This is the most generic of all the approaches and works irrespective of the data mining algorithm that is being used. It typically employs measures like correlation, entropy, mutual information, chi square etc. i.e., analyzing general characteristic of the data to select an optimal feature set. The above are univariate measures i.e. They score each feature individually. There are multivariate scores like Correlation Feature Selection (CFS)[7] and Minimum-redundancy-maximum-relevance (mRMR)[8] feature selection which scores combination of features. Certainly, similar to Wrapper, Filter methods also in the 'multivariate' case employ similar strategies like - Complete Search, Sequential Search and Randomized Search [3] etc.

This is much simpler and faster to build compared to Embedded and Wrapper approaches, as a result this method is more popular to both academicians and industry practitioner.

*Hybrid Approach:* Hybrid approach combines the philosophy of filtering and wrapping approach. So it uses, properties of data distribution (Filter) to prune the space and then use search strategies as in case of Wrapper to find a subset.

*B. Output of Feature selection algorithm*

The feature selection algorithm can work in two ways, we can use a score for each feature based on entropy, correlation, mutual information etc. and then rank them, and then we can select top *n* or some percentage of the attributes. An exhaustive list of options can be found in [9]. Alternatively, we can formulate feature selection as a search problem and select the best feature subset, which is best for the task at hand or some other measure pertaining to the dataset. For the ranking methods we can use simple metrics based on the association between a feature and the target variable (in a supervised setting), we can also use methods like the Correlation Feature Selection (CFS) where the metrics penalize correlation among the features. An optimal subset approach will have typically the below steps as noted [9] :

I. Selection of initial set of features.
II. Generation of next set of features.
III. Evaluation criteria for the feature subsets (How good that particular subset is?).
IV. Stopping Criteria.

Typically with *N* Features the order of the search space is $O(2^N)$, evaluation of all the feature sets is computationally exhaustive. So generally either greedy or heuristic search is applied. Based on the direction of the search it can be either forward, backward or bi-directional.

*C. Univariate, Multivariate, all types of data:*

Univariate methods treat each feature independently, whereas multivariate methods consider interdependence among the features. Multivariate methods are theoretically sounder. However the speed and simplicity of univariate methods have an intuitive appeal. This bifurcation is applicable for filter methods.

The dataset can have all continuous features; can have few discrete features, all categorical variables, both ordinal and nominal and combinations of all of them. Not all metrics cater to all the varieties as an example the most common metrics indicating the strength of the linear relationship between two variables; Pearson product-moment correlation coefficient needs both the variables to be numeric. Many algorithms require the features to be discrete There is an elaborate note in [42], on use of feature discretization and feature selection in conjunction. Many algorithms deal only with discretized data.

*D. Supervised or Unsupervised:*

Generally a problem in the supervised domain is more clearly formulated and is more intuitive than the unsupervised set. Quite naturally, the problem of feature selection is studied in much more detail for supervised tasks than the unsupervised tasks. There has been also some research in semi-supervised domains, where the availability of the labels is much lesser.

*E. Subset Evaluation:*

The evaluation criteria can be broadly classified [3] as independent and dependent criteria respectively. Filter methods generally use an independent methodology based on different measures like distance, information, dependency and consistency depending on characteristics of training and testing dataset. Dependent methods are applicable for wrappers. Wrappers generally use ,classification accuracy as the most commonly used measure for a supervised task. For an unsupervised task, metrics like cluster compactness etc. can be used. As embedded methods also use, particular algorithms, a dependent criterion is more appropriate for embedded methods.

We summarize the above information in Table 1, which will give the practitioner an idea on what can be the different combinations of the approaches. The one that are italicized are less commonly used. One of the multivariate scores that we discuss in section III is variance inflation factor (VIF). Few optimality criteria are Akaike information Criterion (AIC) and Mallow's Cp and Bayesian information criterion (BIC) respectively.

Table 1. Feature selection methods.

| Type of Method | Univariate/ Multivariate | Supervised/ Unsupervised | Type of Output | Evaluation of subset |
|---|---|---|---|---|
| Filter | Univariate | Supervised | Ranking | Independent |
| | *Multivariate* | *Supervised* | *Ranking* | *Independent* |
| | Multivariate | Supervised | Subset | Independent |
| | Univariate | Unsupervised | Ranking | Independent |
| | Multivariate | Unsupervised | Ranking | Independent |
| | Multivariate | Unsupervised | Subset | Independent |
| | Multivariate | Supervised | Subset | Independent |
| Wrapper | Multivariate | Generally Supervised | Subset | Dependent |
| Embedded | Multivariate | Generally Supervised | Subset | Dependent |

The variants are mostly in filter methods, the wrapper method also can be used for clustering with some cluster validity measure.

*F. Research Trends*

We looked at 100+ research papers from 2003 onwards and looked at the keywords of these papers. We

do not claim that these are completely random and representative sample; however 100 + are a fairly large sample. We have searched for 'Feature Selection' in GoogleScholar and picked up keywords from the papers, depending on their availability. We represent the keywords using bag of words model, so association between the keywords are not captured. Also including abstract, body and conclusion would have given more insight. As keyword fairly accurately represents the central concepts and contribution of the work, we have confined this to keywords as of now. Author provided keywords have been used to do a bibliometric analysis [11],[12].

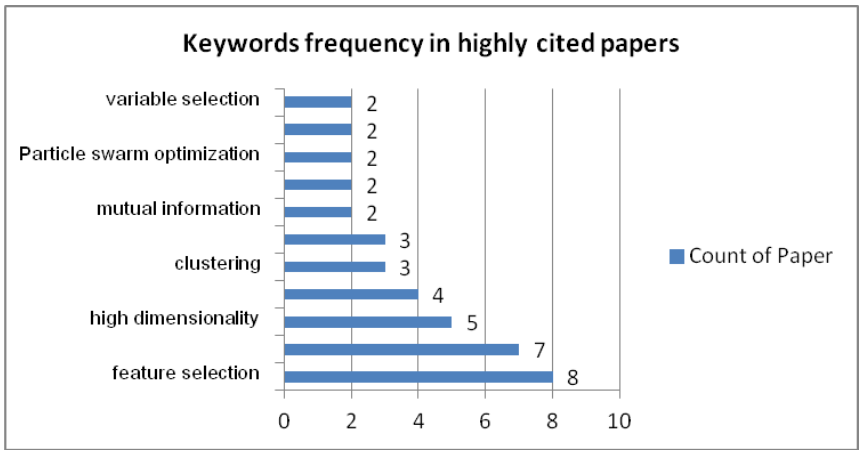


Fig. 1. Keyword counts top cited papers

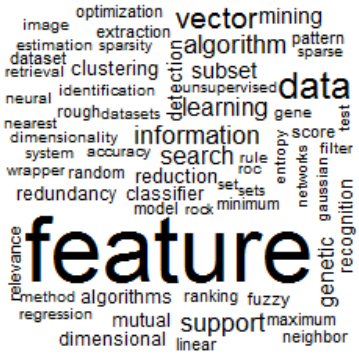


Fig. 2. Word cloud of 100 + papers (Keywords) on feature selection.

The wordcloud is based on unigram and for simplicity. We can figure out importance of support vector machine and genetic algorithms.

Using the citations count to judge the relevance of a paper is in practice for quite some time. There have been many novel methods suggested based on citation in [10] ,based on local and global relation. There can be several other motives for citation of a paper and there are clearly two schools of thoughts one supporting and other against it. However as suggested in [13], relevance of citation count cannot be ruled out as an important parameter to assess a paper's contribution. We picked up few highly cited papers from 2002 onwards (Thresh hold of citation taken as twenty) and we did a key word analysis of the same.

Few things that we notice here are:

i. Supervised Learning is doubly more researched than unsupervised methods.
ii. Genetic Algorithm and Support Vector Machine are the most used methods.
iii. Many papers consider high dimensionality as important notions to be addressed.

In the below table (Table 2) we summarize some of the unique thought processes in few seminal papers with feature selection:

Table 2. Remarks on various surveys and empirical evaluation in feature selection

| Title/Ref | Unique Contribution | Our Remarks |
|---|---|---|
| An Introduction to Variable and Feature Selection [4] | It has a checklist for feature selection task which consider various aspects like domain knowledge, Normalization of data, interdependence of attributes, data quality, stability. An example where presumably redundant variables adds to performance gain. | Near complete coverage from various facets. As noted by the authors, remarks on benchmark and unifying framework are missing. Notes on real life application domains and applicable methods there in is desired. |
| Toward Integrating Feature Selection Algorithms for Classification and Clustering [3] | A categorization framework from designer's point of view based on search strategies (Complete, Sequential, and Random), data mining task (Classification, Clustering) and evaluation criteria (Hybrid, Wrapper, and Filter) of algorithms. A unifying platform based for user based on Knowledge (Purpose, Time, Output and M/N Ratio) and Data (Class Info, Feature Type, Quality, and N/I Ratio). | This is a fundamental contribution to Feature Selection. An evaluation with few datasets would have evaluated how the unifying platform works. Some other attributes like class skew, attribute skew could have been considered. Limited to clustering and classification. |

| Title/Ref | Unique Contribution | Our Remarks |
|---|---|---|
| A review of feature selection techniques in bioinformatics[6] | Thorough analysis of application domains in Bio-informatics with detailed categorization of domains like microarray domain and mass spectrometry. Lists strategies to deal with small sample domains and also available feature selection packages across different tool stack. | Very exhaustive coverage of bioinformatics domain. A guideline to choose with results would have further helped practitioners. |
| Feature selection: An ever evolving frontier in data mining[5] | Highlights various upcoming areas like symbolic and explanation based learning, ultrahigh dimensional data, explanation based feature selection. | While this is an excellent summarization of FSDM10, a comparison of various methods, benchmarks could have been beneficial. |
| Empirical study of feature selection methods based on individual feature evaluation for classification problems[9] | This is a though empirical study based on five measures (Mutual Information, Gain Ratio , Gini index, Releif-F , Relevance and various methods of selecting attributes like Fixed Number (n), Fraction (F) , threshold (t) etc. These methods are used on a total of thirty five datasets with four classifiers (Na ïve Bayes, kNN, C4.5, ANN). | There are useful recommendations, if the experiments were also extended to high dimensional dataset and covered unsupervised problems, the results would have been more useful. |
| Our Review | Analysis on similarity of 4-6 scores. ( Chi Square, Mutual Information, Information Gain , Symmetrical Uncertainty , Linear Correlation, t – statistic)<br>Analysis of commercial adoption ( SAS, SPSS, Oracle Data Miner, Sql Server Analysis Service)<br>Listing of problems solved across business as well as technical domains | We would further want to do many more empirical studies, to arrive at thumb rules, threshold values.<br>We would like to extend to other areas like association, time series. |

## III. A PERSPECTIVE ON COMMERCIAL ADOPTION

We looked at four commercial tools (SAS, SPSS, Oracle Data Miner and SQL Server Analysis Service), among this four, SAS and SPSS are market leaders in this segment.

**SAS:** As noted in [14], the primary methods are as follows apart from the "automatic selection"

**(i) Correlation Based:** This can be used for variable selection, however the threshold is subjective. This can be applied only on numeric data, tests only linear relationship and is univariate.

**(ii) Variable Clustering**: This uses the concept of clustering features based on their similarity. So highly correlated features are kept in one cluster and their correlation with features in other clusters is very low. Ultimately, one feature is selected from each cluster. It used a *R* square ratio between its own cluster and the other clusters.

**(iii) Variance Inflation Factor (VIF):** This is an interesting concept, where one predictor variable is taken as a target and all other predictor variables are used to predict. Typically, variables with, high VIF is eliminated.

$$VIF = \frac{1}{1 - R_i^2}$$

$R_i^2$, denotes Coefficient of determination for the i$^{th}$ variable.

**SPSS:** SPSS uses a set of univariate for eliminating variables. Few of the methods are like:

(i) Maximum percentage. of missing value.

(ii) Maximum Percentage of records in a single category (If majority of the observations/cases fall into one class then that field won't be of much importance)

(iii) Maximum number of categories as a percentage of records (If a field has too many levels or categories, such as some of the categories have very small number of cases then again that filed won't be of much importance).

(iv) Minimum coefficient of variation and

(v) Minimum standard deviation.

Basically feature selection using SPSS Modeler is a three step process: (i) Screening (ii) Ranking and (iii) Selection. The ranking metrics may vary based on the nature of the data domain. In Table 3 we summarize the same-

Table 3. Different metrics/scores as used in SPSS.

| Types of Variable | Metrics |
|---|---|
| All Categorical, target categorical | 1) Pearson Chi Square 2) Likelihood-ratio Chi Square 3) Cramer's V 4 ) Lambda |
| Some Categorical, target categorical | Pearson Chi Square 2) Likelihood-ratio Chi Square |
| Categorical vs continuous | F Statistic |
| Both Continuous | t-Statistic |

**Oracle Data Miner:** As observed in [15], feature selection is done by ranking the attributes. The metric that is used by oracle is Minimum description Length (MDL) for the same. In MDL, the feature selection is treated as a communication problem; a model is built with each predictor variable and the target. MDL penalizes the model for over-fitting or complexity. As observed by [7], MDL is based on operational definition of Ocaam's Razor which states, given a choice between alternatives which have equivalent results, the one which is simpler should get the preference.

**SQL Server Analysis Service:** As noted in [16], features are scored and then a certain % is selected or number is selected. There are various methods available based on the type of the attributes, algorithm that is being used or the parameters of the model. There is a total of scores that are available.

 

**Interestingness score:** This is useful when all the features are non binary continuous numeric data. This is based on entropy and higher the score, higher the importance of the attribute. Interestingness score is given by

Interestingness (Attribute) = - (m - Entropy(Attribute)) * (m - Entropy(Attribute))

Where Central entropy or m means the entropy of the entire feature set.

The other methods used are Shannon's Entropy, Bayesian with K2 Prior and Bayesian Dirichlet Equivalent with Uniform Prior. All these three scoring methods are available for discrete and discretized attributes.

So few things that we observe are:

i. Most of the methods used by the commercial packages are univariate.
ii. The outputs of most of them are based on ranks.
iii. The various scores are based on information theory or statistics.
iv. We also see, there are many of the popular methods , example using genetic algorithm , or a popular score Relief not being available in main stream data mining package.

## IV. Different Scores/Metrics for Feature Selection

The most common scores and metrics that are used for ranking of variables can be broadly classified as either a statistical score or an information theory based score. In Table 4, we summarize the scores, in terms of their key assumptions, applicability for supervised, unsupervised, advantages, limitation and references of papers where it has been used.

Table 4. Different feature selection scoring method

| Metrics | Assumption/ Remarks | Calculation | Supervised/Unsupervised Univariate/Multivariate Numeric/Discrete/Qualitative | Advantage | Limitation |
|---|---|---|---|---|---|
| **Pearson's Correlation Coefficient** | Relations between the variables are linear. | $\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x\,\sigma_y}$ | Both Supervised and unsupervised. Works in univariate setting. Works only with numeric data. | Very simple to interpret and implement. | ✓ Works only on numeric attributes.<br>✓ Can detect linear relationship. |
| **CFS (Correlation based feature selection** | Proposed in [7], this is based on the fact, the most important attributes are correlated with the class and less with each other. Though it says, to be correlation, this is based on any of the 3 measures.<br>✓ Relief<br>✓ Symmetrical Uncertainty<br>✓ MDL | $r_{zc} = \frac{k\bar{r}_{zi}}{\sqrt{k+k*(k-1)*\bar{r}_{ii}}}$ Where $r_{zc}$ indicates worth of a features subset. $r_{zc}$ is the average of correlation between the features and the target. Variable. $r_{ii}$ is the average inter-correlation between the components. The one with highest $r_{zc}$ is selected. | Supervised , Multivariate Works with all type of data. | Simplicity of the theory. | Does not work for unsupervised. To obtain the optimal feature set, we have to perform a search in the feature subspace which may not be required. |
| **t-statistics** | A null hypothesis is set up assuming that the regression coefficient is zero. We then calculate t-statistics based on the observed value of the sample. t–statistics is used to compute p-value which is the probability of the null hypothesis being true , given the data | $t = \frac{(\beta-\beta_0)}{S.E}$ | Applies only for numeric attributes and is a supervised technique. | Theory is simple and robust. | |
| **Information gain / Mutual Information** | Entropy based | *H(Class) + H(Attribute) - H(Class, Attribute)* | Supervised, work with all type of data. | | It is said to be biased towards features with more value. |
| **Gain ratio** | Entropy based | (H(Class) + H(Attribute) - H(Class, Attribute)) / H(Attribute) | Supervised, work with all type of data. | | |

 

| Metrics | Assumption/ Remarks | *Calculation* | Supervised/Unsupervised Univariate/Multivariate Numeric/Discrete/Qualitative | Advantage | Limitation |
|---|---|---|---|---|---|
| **Symmetrical uncertainty** | Entropy based | *2 * (H(Class) + H(Attribute) - H(Class, Attribute)) / (H(Attribute) + H(Class))* | Supervised, work with all type of data. | | Symmetrical uncertainty gets over the limitation of mutual information. |
| **Crammer's V** | Class separation based on features. | $\sqrt{\frac{\chi^2}{N(K-1)}}$ N : Total no of observation K: No of features, or no. of instances whichever is less. $\chi^2$ is given as $\frac{\sum_{k=1}^{c} n_k(\mu_k^j - \mu_j)^2}{\sigma_j^2}$ Where k denotes the class and j denotes the feature. Similar to Fisher Score | Supervised, work with all type of data. | | There is a criticism when this is applied on high dimensional datasets. Works only in supervised setting. |

## V. Short Review of Application Domains

Feature selections have found its application across various domains and have been used to resolve numerous business problems. In Table 5, we summarize few such application areas. For classifying the problems into sector or area, we either use business domains like banking and finance, insurance, medical and bio informatics etc, or some very typical problems like customer relationship management or unstructured data handling, which is kind of a horizontal that cuts across all the business domains. In Table 5 we highlight different areas of application. The objective was to share the wide range of application and problem domains, rather than a narrow one with respect to a particular domain. Citation and relevance also played a role in listing of the domains along with the problems.

Table 5. Listing of problems addressed across application and problem domain through feature selection

| Problem | Application / Business Domain | Reference |
|---|---|---|
| Electricity Price Forecasting | Power Sector | [17] |
| Bankruptcy Prediction<br>Stock market price index prediction<br>predict the trend of stock markets<br>Stock Price Prediction | Banking and Finance Sector | [18],[19],[20],[21] |
| Insurance risk Classification | Insurance | [22] |
| Parkinson's disease ,<br>Medical Diagnosis of Cardio vascular disease,<br>Cancer Diagnosis<br>Prediction of antimicrobial peptides (Natural anti biotic)<br>Breast cancer diagnosis<br>Early detection of the Alzheimer's disease<br>Brain tumor detection | Medical Science, Bioinformatics | [23],[24],[25],[26],[27],[32] |
| Sentiment Analysis in multiple language<br>Text Clustering<br>Text Classification<br>Emotion recognition from speech<br>Spam Filtering | Unstructured data<br>( Image , Audio ,Video, text ) | [28],[29],[30],[31] |
| Software Effort Estimation<br>Software fault prediction | Software Engineering | [33].[34] |
| Classification of power system disturbances | Power Sector | [35] |
| Intrusion Detection | Network Security | [36],[37],[38] |
| Human identification by gait | Computer Vision | [39] |

While we look at the application domains, two domains that stand out in terms of number and variety of application are *unstructured data and bioinformatics and medical application,* the common linkage that can't be missed is high dimensionality of both these domains.

## VI. Experiment setup and Results

In this section, we list down our approach on the experiment, and discuss results. We confine our experiments in feature selection using scoring methods; the methods employed are univariate. We mainly use Chi-square (similar to fisher score) and information theoretic measures (information gain, mutual information, symmetrical uncertainty) for ranking of the features. Additionally when all the variables (predictors and target) are numeric we have used t-statistics and liner correlation for the same. We use Spearman's rank correlation to compare similarity of two methods. For Classification we have used SVM, Decision Tree and Naïve Bayes. For regression, we have used linear models. The datasets are from UCI Machine Learning Repository [40]

The hardware and software used are as follows:-

Processor: Intel® Core™ Duo CPU T6400 @ 2.00 GHZ

RAM: 4 GB

OS: Windows 7 Ultimate SP1

R: Version 2.15.3 [41]

The experiment has four tracks:

A. Track I: We look at high dimensionality domains like text.
B. Track II: We look at small to medium size dataset for Classification.
C. Track III: We look at small to medium size dataset for regression.
D. Track IV: We look at unsupervised tasks like clustering.

**Track I: High Dimensionality**

The dataset is used is as following:

Table 6. Dataset details for Track I.

| Dataset | Features | No. of Instance | No. of Classes |
|---|---|---|---|
| CNAE-9 | 856 | 1080 | 9 |

Both Chi square and information gain selects same set of 83 attributes, below table summarizes the feature selection. Attributes with non zero values have only been selected; the dataset available as a term document matrix and it uses binary weights.

Table 7. # Selected features by different methods

| Dataset | Features | Selecting using Chi.sqaure | Selecting using Information Gain | Spearman rank Correlation |
|---|---|---|---|---|
| CNAE-9 | 856 | 83 | 83 | 99 |

In Table 8 we summarize results from classification with respect to accuracy, 30% data is used as holdout.

Table 8. Text Classification Result

| Dataset | Naïve Bayes | | SVM | | Decision Tree | |
|---|---|---|---|---|---|---|
| | With All Features | With Selected Features | With All Features | With Selected Features | With All Features | With Selected Features |
| CNAE-9 | 53.08% | 59.56% | 92.28% | 87.65% | 35% | 51% |

So what we observe is that for classifiers like SVM, even seemingly redundant attributes can give better result, where as classifiers like Naïve Bayes and decision tree where the classification accuracy is initially low, with feature selection there is improvement. For more, interested readers we would point them to [44], where we have further extended the empirical study on text classification.

**Track II: Small and Medium Size datasets for Classification**

Table 9, describes the various datasets that we have used here. The characteristics of the classes are as follows:

i. Iris, Wine, Seeds are all numeric predictors , with equal distribution.
ii. Glass and Ecoli have high class skew.
iii. Car is a dataset with categorical attributes.

Table 9. Dataset Details for track II.

| Dataset | Features | No. of Instance | No. of Classes |
|---|---|---|---|
| Iris | 4 | 150 | 3 |
| Wine | 13 | 178 | 3 |
| Seeds | 7 | 210 | 3 |
| Glass | 9 | 214 | 7 |
| Ecoli | 7 | 336 | 8 |
| Car | 6 | 1728 | 4 |
| Ionosphere | 34 | 351 | 2 |

In Table 10, we compare the four measures we discussed for a normal classification problem and also highlight similarity of the results from these methods. Iris and car have been marked 'NA' as essentially all the measures concur on ranking, resulting in all paired rank correlations having value of 1.

Table 10. Comparison of feature selection methods based on Spearman's rank correlation

| Dataset | Most Agreeing scores | Least Agreeing Score | Score with highest Correlation | Score with lowest Correlation |
|---|---|---|---|---|
| Iris | NA | NA | NA | NA |

| Dataset | Most Agreeing scores | Least Agreeing Score | Score with highest Correlation | Score with lowest Correlation |
|---|---|---|---|---|
| Wine | Information Gain and Symmetrical Uncertainty | Information Gain and Gain Ratio | Symmetrical Uncertainty | Gain ratio |
| Seeds | Chi.Square and Symmetrical Uncertainty | Information Gain and Gain Ratio | Symmetrical Uncertainty | Information gain |
| Glass | Information Gain and Symmetrical Uncertainty | Chi square with Information Gain and Symmetrical Uncertainty | Symmetrical Uncertainty, Information Gain | Chi Square |
| Ecoli | Chi Square & gain ratio | Chi Square and Information Gain and symmetrical Uncertainty | All other three | Chi Square |
| Car | NA | NA | NA | NA |
| Ionosphere | Information Gain and Chi Square | Chi Square and gain ratio | Symmetrical Uncertainty | Gain Ratio |

In figure three, similarity or average agreeableness of the methods as an average of individual rank correlations. In Figure Four we show all the possible combination of pair wise similarity expressed in terms of rank correlation.

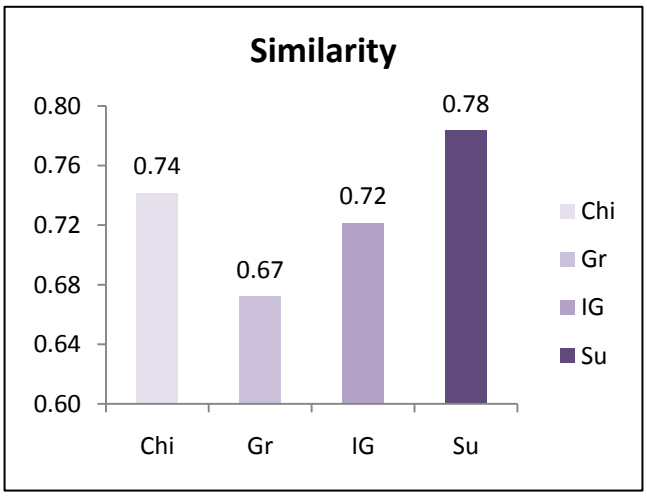


Fig. 3. Summary of Pair wise similarity based on the dataset

**Track III: Regression**

Track III is almost similar to track II. Additionally we compare t – statistics and linear correlation here, as the target attribute in this case is numeric. In Table 11 we summarize the datasets that have been used.

Table 11. Dataset Details for track III.

| Dataset | Features | No. of Instances | No. of Classes |
|---|---|---|---|
| CPU | 6 | 209 | NA |
| Concrete | 8 | 1030 | NA |
| Yachat | 6 | 364 | NA |

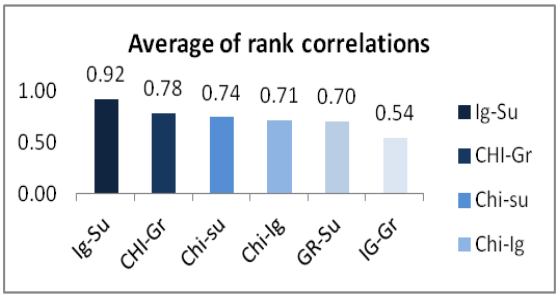


Fig. 4. All Pair-wise similarities. Between different methods

For regression tasks, on contrary to classification, the ranks are widely varying. Some of the correlations are even coming as negative. Example for the CPU dataset, Chi square has a negative correlation of as high as -0.88 with information gain. For CPU, out of six we choose top four (When we fitted the linear model, only for 4 variables p-value was sufficiently small), we used this as a basis of the choice.

Table 12. Dataset Details for track III

| Dataset | Most Agreeing scores | Least Agreeing Score | Score with highest Correlation | Score with lowest Correlation |
|---|---|---|---|---|
| CPU | Gain Ratio and Chi Square | Information Gain and Chi Square | Linear Correlation | Chi Square |
| Concrete | Information gain and symmetric uncertainty | Gain ratio and Chi Square | Information gain and symmetric uncertainty | Gain Ratio |
| Yachat | Information gain and symmetric uncertainty and Chi Square | Gain ratio and Correlation | Information gain and symmetric uncertainty and Chi Square | Gain Ratio |

**Track IV: Clustering**

We do not follow a much formalized method for clustering, which we intend to cover in our future work. The intention is to explore the challenges in clustering. We use supervised datasets; hence for comparing the results we use a supervised measure, *purity* of cluster validity. There is an elaborate framework proposed in[43] which discusses new methods like sparse *k*-means and sparse hierarchical clustering which is tuned for sparse large dimensional dataset and offers much better result

 

than traditional k-means. Through our experiments we actually want to highlight how traditional ways won't work in a small dataset and it needs some additional consideration.

Below is a definition of purity as a supervised cluster validity measure.

Purity: $p_{ij}$ is defined as the probability of a member of cluster $i$ belongs to class $j$, given by $m_{ij}/m_i$, where $m_{ij}$ and $m_i$ are counts as appropriate. Now purity of a cluster $i$ is given by $p_i = max_j\ p_{ij}$ the overall purity is given by $\sum_{i=1}^{k} \frac{m_i}{m} * p_i$

The fundamental assumption is that we will eliminate attributes having very high correlation with one or multiple attributes and preserve the ones which are mostly unrelated.

Table 13. Dataset Details for track IV

| Dataset | Features | No. of Instance | No. of Classes | Type of Attribute |
|---|---|---|---|---|
| Iris | 4 | 150 | 3 | Numeric |
| Seeds | 7 | 210 | 3 | Numeric |

We use the variance inflation factor (VIF) for this purpose, and if we follow the below table we should first eliminate petal length, it is clearly intuitive as close to 97% of the variable is explained by other three , on similar ground sepal width appears to be most novel.

Table 14. VIF Details for IRIS and Seeds datasets.

| Attribute | R Square | VIF |
|---|---|---|
| Petal Length | 0.9674 | 30.67485 |
| Petal Width | 0.9366 | 15.77287 |
| Sepal Width | 0.51 | 2.040816 |
| Sepal Length | 0.8557 | 6.930007 |

| Attribute | R Square | VIF |
|---|---|---|
| Area | 0.9985 | 666.6667 |
| Perimeter | 0.9983 | 588.2353 |
| Compactness | 0.9447 | 18.08318 |
| Lengthofkernel | 0.9791 | 47.84689 |
| Widthofkernel | 0.9904 | 104.1667 |
| asymmetrycoefficient | 0.2634 | 1.357589 |
| lengthofkernelgroove | 0.9152 | 11.79245 |

Table 15. Purity of Iris dataset, with various feature sets

| Dataset | Purity with all attributes | Purity after removing Petal Length | Purity with Petal Length and Petal Width |
|---|---|---|---|
| iris | .89 | .83 | .95 |

So actually, if we have followed the fundamental assumption we would have eliminated petal length and petal height at the starting combination of which actually gives the best cluster validity measure. Similarly with seeds, area seems to be the one feature most explained by other attributes and asymmetry coefficient seems to be most unrelated and hence most novel. Below table gives purity of seeds dataset with all attributes, all attributes except area, all attributes except asymmetric coefficient.

Table 16. Purity of Seeds, with various attribute combinations

| Dataset | Purity with all attributes | Purity after removing area | Purity after removing asymmetric coefficient |
|---|---|---|---|
| Seeds | 0.90 | 0.87 | 0.85 |

Linear Correlation and Information Gain also agree to the same ordering of the attributes. What we observe here is simple, inter correlation is not enough for Clustering and there is a need to work with other characteristics of the data like skew of individual attributes, the overall distribution of the data to name a few.

## VII. Conclusion

Feature selection as a problem is here to stay with the proliferation of variety, volume and velocity of the data. While there have been many surveys and empirical evaluations in this domain, our unique contributions is a review from a commercial side and in depth review of the problems that have been solved across business and technical domains using feature selection. We also present a similarity analysis of the feature selection scores. We observe the application domains for feature selection are vast, with more stress on large dimensional problems like text or bioinformatics. While there have been many types of methods available, in our analysis of commercial software (not open source ones) we found methods used in practice are mostly filter methods using univariate ranking. We conducted our experiment in four tracks. In text classification, SVM had a superior performance in the original feature set itself and there was a drop in the reduced feature set. For other classifiers like Naïve Bayes and Decision Tree , where initial performance was not very good , there was a significant improvement with reduced feature set. A point to mention is , the reduction in case of a text dataset expressed in terms of a term documents matrix, was very significant. We used feature selection on various datasets.

 

We compared next four measures chi square, information gain, gain ratio, symmetrical uncertainty. We used Spearman's rank correlation coefficient to compute similarity between the results. While there was a general agreement between the methods for some of the datasets, the similarity value is as low as 0.14 for some of them. A further analysis in terms of classification accuracy, in terms of highly disagreeing scores can be done. We would recommend symmetrical uncertainty to be the preferred one, with most agreement with other scores and recommend against using gain ratio with least agreement with others. We also looked at few datasets where the target variable is a continuous variable and is a regression problem; the gain ratio remains the one with lowest agreement. For unsupervised learning, we demonstrated with couple of datasets, why still it is a more difficult problem, in spite of some features looking completely redundant , removal of the attributes result in poor cluster quality. There are much more work to be done in unsupervised area, association and time series. We need to use other attributes like class skew, individual attribute skew and any other relevant characteristics into consideration, for feature selection. Also, a study on threshold value of the scores needs to be done to come up with benchmarks.

## REFERENCES


[1] Grantz, John & Reinsel David (2011) Extracting Value from Chaos , IDC I VI EW.

[2] Hilbert, M., & López, P. (2011). The world's technological capacities to store, communicates, and compute information. Science, 332(6025), 60-65.

[3] Huan Liu, Lei Yu (2005) Toward Integrating Feature Selection Algorithms for Classification and Clustering , IEEE Transactions On Knowledge and Data Engineering, VOL. 17, NO. 4, April 2005

[4] Isabelle Guyon , Andr´e Elisseeff (2003) An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157-1182

[5] Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010, June). Feature selection: An ever evolving frontier in data mining. In Proc. The Fourth Workshop on Feature Selection in Data Mining (Vol. 4, pp. 4-13).

[6] Yvan Saeys, In˜aki Inza and Pedro Larran˜aga (2007) A review of feature selection techniques in bioinformatics, Vol. 23 no. 19 2007, pages 2507–2517

[7] Hall, M. A. (1999). Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).

[8] Ding, Chris, and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data." Journal of bioinformatics and computational biology 3.02 (2005): 185-205.

[9] Arauzo-Azofra, Antonio, Jos é Luis Aznarte, and Jos é M. Ben ítez. Empirical study of feature selection methods based on individual feature evaluation for classification problems. Expert Systems with Applications 38.7 (2011): 8170-8177.

[10] Yicong Liang, Qing Li, Tieyun Qian (2011) Finding Relevant Papers Based on Citation Relations , Lecture Notes in Computer Science Volume 6897, 2011, pp 403-414

[11] Xie, S., Zhang, J., & Ho, Y. S. (2008). Assessment of world aerosol research trends by bibliometric analysis. Scientometrics, 77(1), 113-130.

[12] Li, T., Ho, Y. S., & Li, C. Y. (2008). Bibliometric analysis on global Parkinson's disease research trends during 1991–2006. Neuroscience letters, 441(3), 248-252.

[13] Lutz Bornmann and Hans-Dieter Daniel (2008) what do citation counts measure? A review of studies on citing behavior, Journal of Documentation , pp. 45-80

[14] Varun Aggarwal and Sassoon Kosian (2011 ) Feature Selection and Dimension Reduction Techniques in SAS, NESUG 2011

[15] Oracle® Data Mining Concepts 11g Release 1 (2008) http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/feature_extr.htm , B28129-04

[16] Microsoft Developer Network (MSDN), Feature Selection (Data Mining), http://msdn.microsoft.com/en-us/library/ms175382.aspx

[17] Peng, C., Liu, G., & Xiang, L. (2013). Short-term electricity price forecasting using relief-correlation analysis based on feature selection and differential evolution support vector machine. Diangong Jishu Xuebao(Transactions of China Electrotechnical Society), 28(1), 277-284.

[18] Huang, C. L., & Tsai, C. Y. (2009). A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting. Expert Systems with Applications, 36(2), 1529-1539.

[19] Tsai, C. F. (2009). Feature selection in bankruptcy prediction. Knowledge-Based Systems, 22(2), 120-127

[20] Tsai, C. F., & Hsiao, Y. C. (2010). Combining multiple feature selection methods for stock prediction: Union, intersection, and multi-intersection approaches. Decision Support Systems, 50(1), 258-269.

[21] Lee, Ming-Chi. Using support vector machine with a hybrid feature selection method to the stock trend prediction. Expert Systems with Applications 36.8 (2009): 10896-10904.

[22] Duma, M., Twala, B., Nelwamondo, F. V., & Marwala, T. (2012). Partial imputation to improve predictive modelling in insurance risk classification using a hybrid positive selection algorithm and correlation-based feature selection. Current Science(Bangalore), 103(6), 697-705.

[23] Revett, K., Gorunescu, F., & Salem, A. (2009, October). Feature selection in Parkinson's disease: A rough sets approach. In Computer Science and Information Technology, 2009. IMCSIT'09. International Multiconference on (pp. 425-428). IEEE.

[24] Shilaskar, S., & Ghatol, A. (2013). Feature Selection for Medical Diagnosis: Evaluation for Cardiovascular Diseases. Expert Systems with Applications.

[25] Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. (2010). Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics, 26(3), 392-398.

[26] Wang, P., Hu, L., Liu, G., Jiang, N., Chen, X., Xu, J., ... & Chou, K. C. (2011). Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS One, 6(4), e18476.

[27] Chaves, R., Ram írez, J., G órriz, J. M., L ópez, M., Salas-Gonzalez, D., Alvarez, I., & Segovia, F. (2009). SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting. Neuroscience Letters, 461(3), 293-297

[28] Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for

opinion classification in Web forums. ACM Transactions on Information Systems (TOIS), 26(3), 12.
[29] Li, Y., Luo, C., & Chung, S. M. (2008). Text clustering with feature selection by using statistical data. Knowledge and Data Engineering, IEEE Transactions on, 20(5), 641-652.
[30] Chen, J., Huang, H., Tian, S., & Qu, Y. (2009). Feature selection for text classification with Naïve Bayes. Expert Systems with Applications, 36(3), 5432-5435.
[31] Rong, J., Li, G., & Chen, Y. P. P. (2009). Acoustic feature selection for automatic emotion recognition from speech. Information processing & management, 45(3), 315-328.
[32] Meher, Jayakishan, et al. "Cascaded Factor Analysis and Wavelet Transform Method for Tumor Classification Using Gene Expression Data." International Journal of Information Technology & Computer Science 4.9 (2012).
[33] Oliveira, A. L., Braga, P. L., Lima, R. M., & Cornélio, M. L. (2010). GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation. information and Software Technology, 52(11), 1155-1166.
[34] Catal, C., & Diri, B. (2009). Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. *Information Sciences*, *179*(8), 1040-1058.
[35] Erişti, H., Uçar, A., & Demir, Y. (2010). Wavelet-based feature extraction and selection for classification of power system disturbances using support vector machines. Electric power systems research, 80(7), 743-752.
[36] Tsang, C. H., Kwong, S., & Wang, H. (2007). Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection. Pattern Recognition, 40(9), 2373-2391
[37] Nguyen, H., Franke, K., & Petrovic, S. (2010, February). Improving effectiveness of intrusion detection by correlation feature selection. In Availability, Reliability, and Security, 2010. ARES'10 International Conference on (pp. 17-24). IEEE.
[38] Amiri, F., Rezaei Yousefi, M., Lucas, C., Shakery, A., & Yazdani, N. (2011). Mutual information-based feature selection for intrusion detection systems. Journal of Network and Computer Applications, 34(4), 1184-1199.
[39] Bashir, K., Xiang, T., & Gong, S. (2008, March). Feature selection on gait energy image for human identification. In Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (pp. 985-988). IEEE.
[40] Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
[41] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org/.
[42] Ferreira, A. J., & Figueiredo, M. A. (2012). An unsupervised approach to feature discretization and selection. Pattern Recognition, 45(9), 3048-3060
[43] Witten, D. M., & Tibshirani, R. (2010). A framework for feature selection in clustering. Journal of the American Statistical Association, 105(490).
[44] Sarkar, Subhajit Dey, and Saptarsi Goswami. "Empirical Study on Filter based Feature Selection Methods for Text Classification." International Journal of Computer Applications 81 (2013).

## Authors' Profiles

**Saptarsi Goswami:** He is an Assistant professor at Institute of Engineering and Management in Computer Science Department, India and a Research Scholar at A.K.Choudhury School of Information Technology, University of Calcutta. He has 10 + Years of working experience in IT industry. His area of interest is feature selection, outlier detection, mining unstructured data etc.

**Amlan Chakrabarti:** He is at present an Associate Professor and HoD at the A.K.Choudhury School of Information Technology, University of Calcutta. He has done his Doctoral research on Quantum Computing and related VLSI design at Indian Statistical Institute, Kolkata, 2004-2008. He was a Post-Doctoral fellow at the School of Engineering, Princeton University, USA during 2011-2012. He is the recipient of BOYSCAST fellowship award in the area of Engineering Science from the Department of Science and Technology Govt. of India in 2011.He has held Visiting Scientist position at the GSI Helmohltz research laboratory Germany and Department of Computer Science and Engineering at the New York State University at Buffalo, U.S.A. during Sept-Oct., 2007. He has published around 50 research papers in referred journals and conferences. He is a Sr. Member of IEEE and life member of Computer Society of India. He has been the reviewer of IEEE Transactions on Computers, IET Computers & Digital Techniques, Elsevier Simulation Modeling Practice and Theory, Springer Journal of Electronic Testing: Theory and Applications. His research interests are: Quantum Computing, VLSI design, Embedded System Design, Video and Image Processing Algorithms and pattern recognition.