# A study on Feature Selection Methods in Medical Decision Support Systems

Rahul Samant,
*SVKM'S NMIMS, Shirpur Campus, India;*

Srikantha Rao,
*TIMSCDR, Mumbai University, Kandivali, Mumbai, India,*

## Abstract

*Clinical databases often consist of a large number of disease markers. For clinical data analysis, some disease markers are not helpful and sometimes even have negative effects. Therefore, applying feature selection is necessary as it can remove those unimportant disease markers. It also increases the effectiveness of Medical Decision support system by effectively decreasing learning time of the system. We evaluated three different feature selection methods, such as Principle Component Analysis (PCA), Factor Analysis ( FA), and Attribute Ranking(AR) method. Finally, the promising performance of PCA was validated through a set of experiments on a dataset using Naïve Bayesian (NB) classifier and K-nearest neighbor (KNN) classifier.*

## 1. Introduction

Machine learning techniques have been widely used to help the medical experts in analyzing medical data [1]. Generally large scale medical databases are having large number of attributes or dimensions. This large data dimensionality can badly influence many aspects of analysis process. It can increase learning system's time on both training and runtime phases. Meanwhile, it may cause the "curse of dimensionality" problem. To handle the high dimension medical data, feature reduction is an important technique [4-6]. Researchers and practitioners realize that in order to use data mining tools effectively data preprocessing is essential to successful data mining [13]. The idea of feature reduction is to use fewer dimensions of data to represent original data. Note that although the number of dimensions is reduced, the discriminative capability should not be hampered. There are many benefits with feature reduction. For example, it can avoid over-fitting, reduce data analysis complexity and improve data analysis performance. Generally, feature reduction can be divided into two categories [7]: feature extraction and feature selection. In feature extraction, the original feature space is mapped into a lower dimensional one. Therefore, the features are totally new and different with original features. The popular feature extraction methods include principal component analysis (PCA) and independent component analysis (ICA) [8]. Through feature extraction, although a much smaller dimension is obtained, this usually requires high computational overhead. In addition, since the features are new, it is hard to interpret by human. For example, the disease markers are the popular features for medical data. Through feature extraction, these disease markers will be projected to another space. In that case, the medical doctor cannot interpret what is the meaning of the new features. Feature selection will choose partial features from original feature spaces according to a specified evaluation function. Usually this evaluate function evaluates the discrimination capability of each feature. Unlike feature extraction, feature selection does not generate any new features. This will not make any difficulties for human understanding the meaning of features. Due to this advantage, feature selection is more widely used for medical data analysis compared to feature extraction. Mainly there are three kinds of feature selection methods [9]: wrapper, filter, and embedded methods. In the embedded method, feature selection is one part of the learning algorithm. C4. 5 is one of the typical embedded methods [10]. Wrappers [11] evaluate the prediction performances of features by employing a specified learning algorithm. Based on the learning algorithm, the feature which gives higher classification accuracy will be selected. In general, the quality of features selected by wrappers is good as classification accuracy is directly taken into account. But meanwhile, wrapper also has some limitations. Firstly, the features selected by wrapper depend on the learning algorithm. They might be not suitable for other learning algorithms. Secondly, wrapper is time-consuming and often intractable for large-scale problems. Different with wrappers, filters do not employ any specified learning algorithm. Instead, it identifies a subset of features according to some evaluation criterions. Different evaluate criterions are used by various filters. Filter is independent with learning algorithm and computational efficient. Therefore, it has been widely applied on medical data [12].

## 2. Methods for Feature Selection

In this section of the paper, we briefly describe the three attribute selection methods used in our study. The excellent summary of feature selection methods is provided in the book “Data Preparation for Data Mining” by D. Pyle.[13]
.

### 2.1 Principle Component Analysis (PCA)

Principal component analysis (PCA) is the best, in the mean-square error sense, linear dimension reduction technique. Being based on the covariance matrix of the variables, it is a second-order method. The basic procedure is as follows:

a. The input data is normalized to ensure that attributes with large domains will not dominate attributes with smaller domain.
b. PCA computes K orthogonal vectors that provide a basis for the normalized input data. These vectors are referred to as principle components.
c. The principle components are sorted in order of decreasing ‘significance’ or strength. They essentially serve as a new set of axes for data, providing important information about variance.
d. Because the components are sorted according to decreasing order of ‘significance’ the size of the data can be reduced by eliminating the weaker components. Using the strongest components it should be possible to reconstruct a good approximation of the original data.

### 2.2 Factor Analysis (FA)

Like PCA, factor analysis (FA) is also a linear method, based on the second-order data summaries. FA assumes that the measured variables depend on some unknown, and often not measurable, common factors. For examples, for many psychiatric data is not possible to measure a certain factor of interest directly (such as “intelligence”); however it is possible to measure other parameters such as various test scores of individuals to reflect the “intelligence" factor. The goal of FA is to uncover such relations, and thus can be used to reduce the dimension of datasets following the factor model.

### 2.3 Median Imputation (MDI)

Since the mean is affected by the presence of outliers it seems natural to use the median instead just to assure robustness. In this case the missing data for a given feature is replaced by the median of all known values of that attribute in the class where the instance with the missing feature belongs. This method is also a recommended choice when the distribution of the values of a given feature is skewed.

### 2.3 Attribute Ranker (AR)

In this algorithm evaluation a subset of 1 or more features is given a goodness score by the evaluator (SubsetEvaluator); in attribute evaluation the evaluator (AttributeEvaluator) gives each individual attribute a goodness score. When a search algorithm is paired with a SubsetEvaluator it explores the space of possible subsets and returns the best one with respect to the evaluation. In the case of the later, a ranked list of attributes is produced by pairing the attribute evaluator with a special "search" called the Ranker. For ranked lists of attributes you can easily specify that you want to retain the top M ranked attributes (or alternatively, set a threshold on the goodness score by which to discard some of the ranked attributes). It is still possible to specify that you want the best M attributes when using a subset evaluator by pairing it with the GreedyStepwise search method and turning on the option to produce a ranked list. This works by forcing the search to the far side of the search space by continuing to add attributes to the best subset even if it decreases the overall goodness (it still adds the "best" attribute at each stage - i.e. the one that that decreases the goodness the least). The order that attributes are added forms the ranking.

## 3. Experiments

### 3.1 Dataset Acquisition

The database used for analysis in this paper has been compiled as a part of an earlier study entitled Early Detection Project (EDP) conducted at the Hemorheology Laboratory of the erstwhile Inter-Disciplinary Programme in Biomedical Engineering at the School (now department) of Biosciences and Bioengineering, Indian Institute of Technology Bombay (IITB), Mumbai, India. Spanning over a period from Jan.1990 to Apr. 1996, it consists of 968 records, each with 30 parameters, which encapsulate the biochemical, hemorheological and clinical status of the individuals visiting Hospital for routine checkups or treatment of common ailments. We note that the Hemorheology Laboratory has pioneered the research in the field of Clinical Hemorheology by conducting the baseline hemorheological studies in the Indian population and correlating various hemorheological parameters with several disease conditions.

### 3.2 Profile of the sample

In all, 30 parameters have been noted for each respondent. They include age, gender, habit (smoking , alcohol consumption), blood groups, disease state, health indicators (e.g.; pulse, systolic blood pressure (BP1), diastolic blood pressure (BP2) ) and biochemical parameter like Serum

Proteins (SP), Serum Albumin (SALB), Serum Fibrinogen (SFIB), Hematocrit (HCT), Erythrocyte Sedimentation Rate (ESR), Serum Cholesterol (SC), Serum Triglycerides (STG), Hemoglobin (HB), Platelet Aggregation (PLA), along with various hemorheological (HR) parameters (e.g.; Whole Blood Viscosity – WBV- measured over eight different shear rates, Plasma Viscosity- PV- measured over three different shear rates, using a Contraves 30 viscometer, and Red Cell Aggregation - RCA).

The database covers a very wide age-range (14 - 83 years), although majority of the respondents are closer to 40 years of age (average age 41.67 years). The female component of the database is comparable to the male component in most parameters like age, blood group distribution and biochemical values. The average BP of the entire database (127.27 /83.89 mm of Hg) and also that of its male (127.97/85.14 mm of Hg ) and female (125.46/82.36 mm of Hg ) components are close to the normal value of 120/80 mm of Hg, reported in the literature, signifying preponderance of normal controls in the sample. About 16% of male subjects indulge in smoking, alcohol consumption or both, while the corresponding value for females is 1%. The distribution of blood groups in the database is consistent with other reports related to Indian population. The incidence of HT among the database studied is found to be higher in males as compared to females (21.45% and 13.10% resp.); while the corresponding figures for Diabetes Mellitus (DM) in these sexes are 15.24% and 12.46 % resp. Most biochemical parameters are found to lie within the normal range. Plasma Viscosity (PV) shows 100% variation between maximum and minimum reported values (1.02 to 2.02cp), both in males and females. The average plasma viscosity in females is slightly higher than that reported in males (1.40 cp against 1.385 cp), the difference is not statistically significant. On the contrary, the Whole Blood Viscosity at high shear rate (WBVh) is significantly higher in males than in females (5.47 cp against 4.48 cp, p<0.01). The higher PV in females may be attributed to a higher percentage of females reporting elevated PF values and is consistent with earlier reported studies.

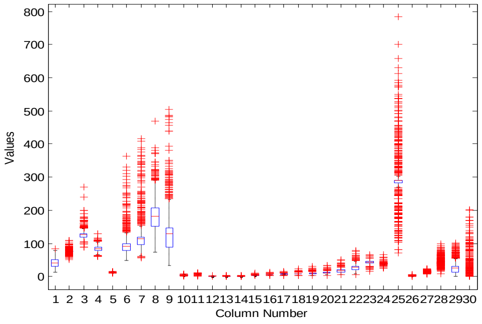

**Fig. 1. The dataset visualization**

### 3.1 Data preprocessing

The database was having missing data values at random. The data records having missing values were dropped entirely from the database and a cleaned dataset was prepared. We applied Principle Component Analysis, Factor analysis and Attribute Ranking techniques for feature selection. The score given to each attribute by respective feature selection method results are tabulated in Table 1.

**Table 1. Score of selected features**

| Feature | PCA | Factor analysis | Attribute Ranker |
|---|---|---|---|
| AGE | 37.3121 | 0.9989 | 0.3585 |
| BP1 | 26.1401 | 0.9776 | 0.5201 |
| BP2 | 15.9286 | 0.9620 | 1.0605 |
| HB | 10.5788 | 0.9876 | 0 |
| BSF | 3.3159 | 0.9987 | 0.9392 |
| BSP | 2.1825 | 0.9995 | 0.7714 |
| SC | 1.6910 | 0.9993 | 0.2136 |
| STG | 0.9856 | 0.9778 | 0.1919 |
| SALB | 0.6781 | 0.9924 | -0.1789 |
| SP | 0.5222 | 0.9923 | 0.3984 |
| CPV1 | 0.3823 | 0.9203 | 0.1644 |
| CPV2 | 0.1479 | 0.8966 | -0.1247 |
| CPV3 | 0.0714 | 0.8874 | 0.1743 |
| CB1 | 0.0195 | 0.1507 | 0.3943 |
| CB2 | 0.0182 | 0.0982 | 0.3586 |
| CB3 | 0.0116 | 0.0383 | 0.4028 |
| CB4 | 0.0044 | 0.0050 | 0.576 |
| CB5 | 0.0034 | 0.0207 | -0.0636 |
| CB6 | 0.0018 | 0.0523 | -0.4214 |
| CB7 | 0.0017 | 0.1113 | -0.1953 |
| CB8 | 0.0014 | 0.2277 | 0.1153 |
| HCT | 0.0005 | 0.5149 | 0.1296 |
| RHCT | 0.0004 | 0.9963 | 0 |
| SF | 0.0002 | 0.9992 | -0.4776 |
| RG | 0.0002 | 0.9838 | 0.0238 |
| PA | 0.0001 | 0.9974 | -0.1153 |
| PLA | 0.0001 | 0.9999 | -0.0025 |
| ESR4 | 0.0000 | 0.9610 | -0.3187 |
| ESR_COR | 0.0000 | 0.9902 | 0.1387 |

## 4. Results and Discussion

In the experiments, we divided the dataset into two subsets. First subset consists of data related to hypertensive and normal patients. Second subset contains mixed population data about diabetic and normal patients. We used Naïve Bayesian (NB) classifier and K-nearest neighbor (KNN) classifier to evaluate the effectiveness of the three attribute selection methods. For classification models, we consider four different feature set of the datasets. In the first attempt we use all feature set for building classification model. Subsequently we considered selected features by PCA, FA and AR methods to build the classification models. MatLab 2007a tool was used to code and test these classification models. Classifier accuracy was used as a measure to determine the effectiveness of feature selection methods. As shown PCA resulted in best classification accuracy for both, Naïve Bayesian as well as KNN classifier. When we use PCA the classification accuracy for Naïve Bayesian Classifier increased from 62.34% for dataset-1 to 84.33% and for dataset-2, accuracy improved from 50.34% to 79.21%, compared to consideration of all features for classification. Similarly for KNN classifier the accuracy improved from 59.13% to 80.25% for dataset-1 and for dataset-2 it increased from 52.17% to 82.58%, compared to consideration of all features for classification. Other feature selection methods also show promising increase of accuracy compared to accuracy displayed by considering all the attributes (features) for building classification model. But PCA recorded best results.

**Table 1: Datasets used in the study**

| # | Dataset | Features | Instances |
|---|---|---|---|
| 1 | Diabetes dataset | 30 | 235 |
| 2 | Hypertension dataset | 30 | 338 |

**Table 2: Experimental results**

| Classifier | Data set | % Classifier accuracy | | | |
|---|---|---|---|---|---|
| | | All features | PCA | FA | AR |
| Naïve Bayesian classifier | 1 | 62.34 | 84.33 | 80.21 | 74.43 |
| | 2 | 50.34 | 79.21 | 75.38 | 65.69 |
| KNN classifier | 1 | 59.13 | 80.25 | 76.83 | 68.25 |
| | 2 | 52.17 | 82.58 | 74.32 | 66.34 |

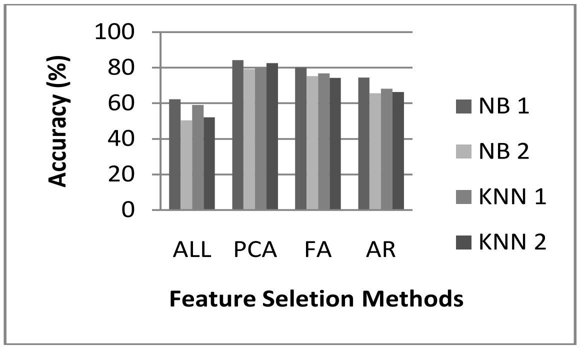

**Fig. 2. Effect of attribute selection on classifier accuracy**

## 4. Conclusion

Medical data usually consists of a large number of disease markers, it is hard to analyze by humans. In building medical decision support systems a small subset of relevant disease markers or symptoms (features) are needed to show higher accuracy and lower learning time. Generally a large amount of diagnosed samples are required to achieve the good feature selection performance. The feature selection methods definitely improve the classification accuracy. In our study all the three methods viz. PCA, FA and AR shows promising performance to improve the accuracy of the classification models. The performance of PCA was the best. Therefore, future work will investigate the using of undiagnosed samples for other advanced feature selection methods. Moreover, feature selection is a kind of data preprocessing technique for medical data. In addition to it, there exist other preprocessing techniques such as noisy medical data detection yet to be explored.

## 5. References

[1] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine* 23 (1), 2001, pp. 89-109.

[2] T. T. Tang, G. Zheng, Y. L. Huang, G. F. Shu and P. T. Wang, "A comparative study of medical data classification methods based on decision tree and system reconstruction analysis," *Journal of EMS* 4, 2005, pp. 102-108.

[3] W. Wajs, P. Wais, M. Swiecicki, and H. Wojtowicz, "Artificial immune system for medical data classification," in: Proc. of 2005 *International Conference on Computational Science*, 2005, pp. 810-812.

[4] T. H. Cheng, C. P. Wei, V. S. Tseng, "Feature selection for medical data mining: omparisons of expert judgment and automatic approaches," in: Proc. of the 19th *IEEE Symposium on Computer based Medical Systems*, 2006, pp. 165-I 70.

[5] K. Polat and S. Gunes, "A new feature selection method on classification of medical datasets: Kernel F-score feature selection,"*Expert Systems with Applications* 36 (7), 2009, pp. 10367-10373.

[6] L. Chuang, H. Chang, C. Tu, and C. Yang, "Improved binary PSO for feature selection using gene expression data," *Computational Biology and Chemistry* 32 (1), 2008, pp. 29-38.

[7] M. Raymer, W. Punch, E. Goodman, L. Kuhn, and A. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, 2000, pp. 164-171,

[8] A. Hyvarinen, E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks* 13 (4-5), 2000, pp. 411-430.

[9] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," 1. *Mach. Learn. Res*. 3,2003, pp. 1157-1182.

[10] L. Breiman, 1. H. Friedman, R. A. Olshen, and C. 1. Stone, Classification and Regression Trees, Wadsworth and Brooks, 1984.

[11] R. Kohavi and G. John, "Wrappers for feature selection," *Artificial Intelligence* 97 (1-2) ,1997, pp. 273-324.

[12] K. Kira and L. Rendell, "A practical approach to feature selection," In Proc. *International Conference on Machine Learning*, 1992, pp. 368-377.

[13] D. Pyle. “*Data Preparation for Data Mining*”, Morgan Kaufmann Publishers, 1999.