International Journal of Electrical, Electronics and Computer Engineering **5**(1): 27-32(2016) **ISSN No. (Online): 2277-2626**

Feature Selection/Reduction and Classification in large Datasets using Data Mining Concepts: A Review

***Sankia Mishra**[*] **and Prof. Dr. Vineet Richhariya**[**]*

[*]M. Tech. Research Scholar, Department of Computer Science & Engineering, Laxmi Narayan College of Technology, Bhopal, INDIA

*** HOD Department of Computer Science & Engineering, Laxmi Narayan College of Technology, Bhopal, INDIA*

(Corresponding author: Sankia Mishra)

(Received 10 February, 2016 Accepted 17 March, 2016)

(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Feature selection and reduction is a tedious task in the field of data mining from the huge database. It is not necessary that the information which is available to us all is crucial and efficient. So we must eliminate the redundant and irrelevant feature from dataset. The feature selection and reduction is performed by using data mining classification algorithm such as clustering, K-nearest neighbor classifier etc. In this, we use wine dataset from UCI Repository of Machine learning databases. This paper presents the literature study of the formerly work done in the feature selection and reduction from the huge database. We also presents different data mining classification algorithm which helps in feature reduction.

Keywords: Classification, Feature selection, Machine learning, Reduction.

I. INTRODUCTION

As the time passes, there is incredible development in the use of internet for sharing personal data and resources but sometimes its extensive use becomes more vital for us. The internet also fashioned copious ways to conciliate the immovability and protection of the systems connected to it.

The unrelated input features will persuade greater computational cost and lead to over fitting. Feature reduction is the most challenging task from the outsized database. This is performed to eliminate the redundant features from the original data which improves classifier performance. Feature selection [1] has been exceptionally active area of research and improvement of clustering and pattern recognition in machine learning, data mining, statistics, biology etc. It has been observed that: a) The bulky number of features is not informative because irrelevancy or redundancy with respect to the cluster (class) perception; b) learning can be achieved more resourcefully and effectively with just appropriate and non-redundant features. On the other hand, finding best possible feature subset is usually very much intricate and many problems related to feature selection. Selecting good features is a crucial activity and requires extensive domain knowledge.

There are numerous feature selection systems that are created either to choose the elements or concentrate highlights. A developmental methodology for highlight determination is proposed which depends on numerical crossing point rule. Data mining is fraction of the knowledge discovery in databases (KDD) procedure and it is the process of analyzing collected data to discover patterns or correlations. The KDD process can be seen as five steps: Data selection, data pre-processing, data reduction, data mining and explanation/valuation. Figure 1 shows the steps involved to process the data in data mining. Clustering is a kind of classification method for data with unknown allocation; the purpose is to discover the structure concealed in data and as much as feasible, to bring together the data with similar nature to the similar cluster according to some measure of likeness degree. Most clustering algorithms do not rely on suppositions common to conventional statistical methods, i.e. underlying statistical distribution of data and consequently they are useful in situations where little former knowledge exists. The impending of clustering algorithms to divulge the underlying structures in data can be employed in a broad range of applications, together with classification, image processing, modeling and recognition [2, 3].

The feature selection process is performed on wine dataset which is mainly used for the research work. These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines. In this paper, we present the literature study about the feature selection and reduction performed by different researchers using classification techniques of data mining. The organization of remaining section of the paper is done as follows: Section II presents the overview about the feature selection process. In section III, we discuss methods of feature selection and reduction. In section IV discuss about data mining and its classification algorithm. Section V describes the overall conclusion of the paper and its future work.

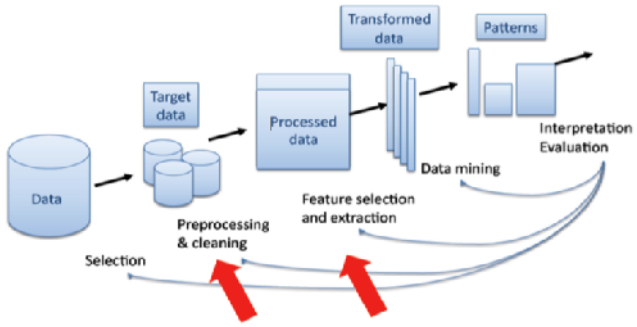

Fig. 1. Steps involved in the process of Data Mining.

II. FEATURE SELECTION PROCEDURE

The four key steps of a Feature selection process are feature subset generation, subset evaluation, stopping criterion and result validation. The feature subset generation is a heuristic search process which results in the selection of a candidate subset for evaluation. It uses searching strategies like complete, sequential and random search to generate subsets of features. In [4] stated that these searching strategies are based on stepwise addition or deletion of features. The goodness of the generated subset is evaluated using an evaluation criterion. If the newly generated subset is better than the previous subset, it replaces the previous subset with the best subset. These two processes are repeated until the stopping criterion is reached. The final best feature subset is then validated by prior knowledge or using different tests. Fig.2 illustrates the feature selection process. The main objectives of feature selection can be as follows:

-Improving the performance of learning algorithm.
-Reducing the storage requirement for data set.
-Enhancing data understanding and helping to visualize it.
-Detecting noises and outliers in the data set.

Feature selection is an innovative area of research in pattern recognition, machine learning, and data mining and is widely applied to many fields such as text categorization, image retrieval, customer relationship management, intrusion detection.

Generation = select feature subset candidate.
Evaluation = compute relevancy value of the subset.

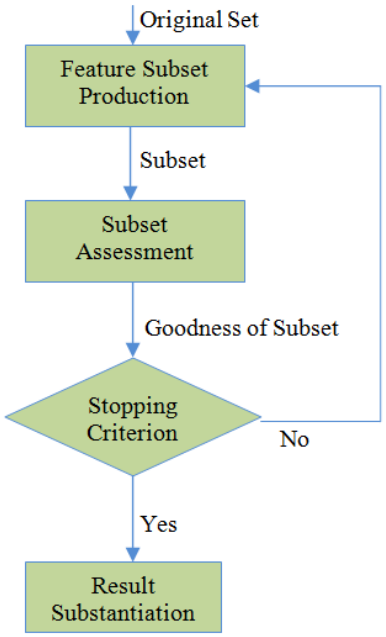

Fig. 2. Feature selection Procedure.

Stopping criterion = determine whether subset is relevant.
Validation = verify subset validity
Feature extraction, an alternative approach of data reduction, is a special type of dimensionality reduction method to create a set of new reduced features based on some transformation function.

III. FFEATURE SELECTION AND REDUCTION METHOD

Feature selection is the process of selecting a subset of features from the entire collection of available features of the dataset. Thus for feature selection, no preprocessing is required as in case of feature extraction. Usually the objective of feature selection is to select a subset of features for data mining or machine learning applications [5]. Feature selection [5, 6, 7] can be achieved by using supervised and unsupervised methods. The process of Feature selection is based mainly on three approaches viz. filter, wrapper [8] and embedded (Fig. 3).

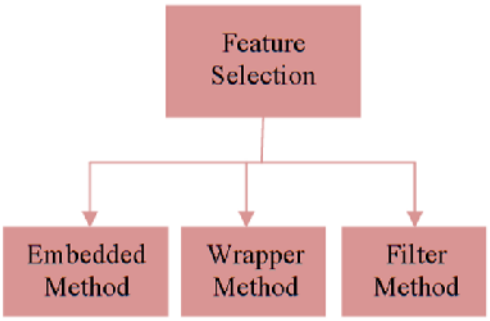

Fig. 3. Feature selection techniques.

A. Filter Method [9]

The filter approach incorporates an independent measure for evaluating features subsets without involving a learning algorithm. This approach is efficient and fast to compute (computationally efficient). However, filter methods can miss features that are not useful by themselves but can be very useful when combined with others. The graphical representation of the filter model is shown in Figure 4.

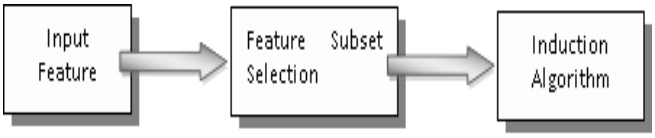

Fig. 4. Model of filter method.

B. Wrapper Method [10]

Wrapper Method requires a predetermined algorithm to determine the best feature subset. Predictive accuracy of the algorithm is used for evaluation. This method guarantees better results, but it is computationally expensive for large dataset. For this reason, the Wrapper method is not usually preferred.

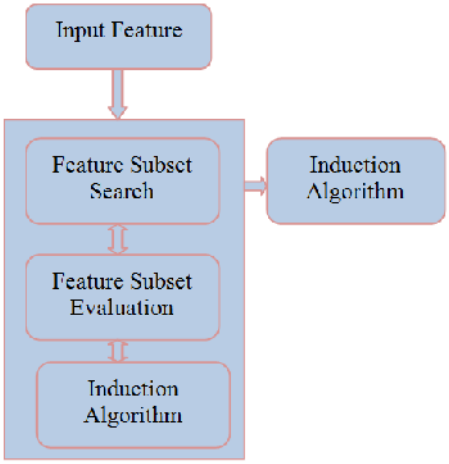

Fig. 5. Model of Wrapper method.

C. Embedded Model

Embedded Models embedding feature selection with classifier construction have the advantages of (1) wrapper models - they include the interaction with the classification model and (2) filter models - they are far less computationally intensive than wrapper methods [11, 12, 13]. There are three types of embedded methods. The first are pruning methods that first utilizing all features to train a model and then attempt to eliminate some features by setting the corresponding coefficients to 0, while maintaining model performance such as recursive feature elimination using support vector machine (SVM) [14]. The second are models with a build-in mechanism for feature selection such as ID3 [16] and C4.5 [15]. The third are regularization models with objective functions that minimize fitting errors and in the mean time force the coefficients to be small or to be exact zero. Features with coefficients that are close to 0 are then eliminated [12]. Due to good performance, regularization models attract increasing attention.

IV. DATA MININING ALGORITHM

Data Mining (The analysis step of the knowledge discovery in data base) a powerful new technology improved and so fast grown. It is a technology used with great potential to help business and companies focus on the most important information of the data that they have to collect to find out their customer's behaviors. Intelligent methods are applied in order to extracting data pattern, by many stages like" data selection, cleaning, data integration, transformation and pattern extraction". Many methods are used for extraction data like" Classification, Regression, Clustering, Rule generation, Discovering, association Rule…etc. each has its own and different algorithms to attempt to fit a model to the data. Algorithm is a set of rules that must be followed when solving a specific problem (it is a finite sequence of computational steps that transform the given input to an output for a given problem). The problem can be a machine. Classification techniques in data mining are capable of processing a large amount of data. It can predict categorical class labels and classifies data based on training set and class labels and hence can be used for classifying newly available data. Thus it can be outlined as an inevitable part of data mining and is gaining more popularity [17]. There are different classification algorithms such as KNN Classifier, decision tree, Bayesian tree, PSO etc. which are explained below in detail.

A. Particle Swarm Optimization

PSO was coined by James Kennedy and Russell Eberhart in 1995 after being stimulated by the social communication behavior of a group of birds by biologist Frank Heppner. It is more or less equivalent to evolution inspired problem solving methods like genetic algorithms. Many enhanced PSO algorithms were implemented by different scholars.

Nowadays PSO is one of the most challenging subject in the field of natural computing and been implemented in organizations successfully. PSO is a very strong stochastic optimization approach. It depends on the movement and intelligence of swarms. PSO [18][19] is a heuristic search methodology, applicable in the subject of social communication to problem solution. The PSO algorithm manages a population of particles, where each and every particle constitutes a prospective solution to an operation research problem. In PSO, each and every particle travels through the solution space. It computes its strength depending on its current position and the complete population's best position. The most appropriate solution is called as the fitness function. The fitness function keeps on changing with respect to various goals. To obtain a good solution a number of agents are used in PSO that made up of a swarm flying in the region of the exploration space. Each and every agent is considered as a spot in a N-dimensional space. It tunes its "moving" according to its own moving practice as well as the moving practice of the remaining particles. Each particle manages its end points in the solution space according to the fitness function that was gained till now by that particle. This value is represented as personal best, pbest. The next best value, tracked by the PSO is nothing but the best value got so far by any particle in the nearby region of that particle. This value is represented as gbest.

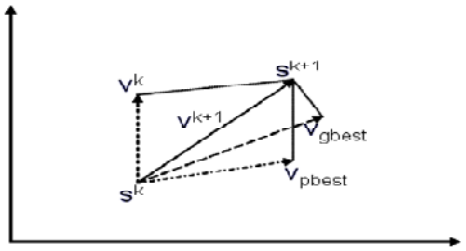

Fig. 6. Change in a searching point by PSO.

B. K-Nearest Neighbor Classifier [20]

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample.

"Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, X=(x1, x2, xn) and Y=(y1,y2,….,yn) is

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

The neighbors are taken from a set of objects for which the correct classification (or, in the case of regression, the value
of the property) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. In order to identify neighbors, the objects are represented by position vectors in a multidimensional feature space. It is usual to use the Euclidian distance, though other distance measures, such as the Manhattan distance could in principle be used instead. The k-nearest neighbor algorithm is sensitive to the local structure of the data.

C. Bayesian Networks [21]

A Bayesian network or probabilistic directed acyclic graphical model is a type of statistical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms.

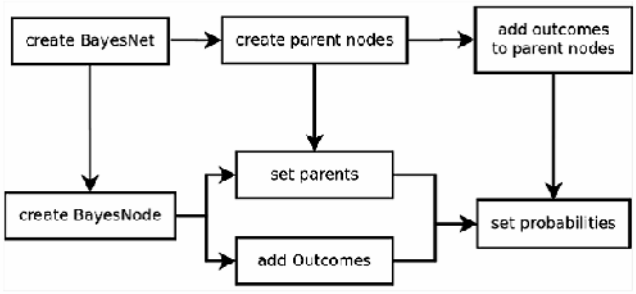

Fig. 7. Bayesian Network.

The Bayes Network is a container for Bayes Nodes, which represent the random variables of the probability distribution of modeling. A Bayes Node has outcomes, parents, and a conditional probability. It is important to set the probabilities last, after setting the outcomes of the parent nodes.

Bayesian network can take the following forms:

1. Declaring that a node is a root node, i.e., it has no parents.
2. Declaring that a node is a leaf node, i.e., it has no children.
3. Declaring that a node is a direct cause or direct effect of another node.
4. Declaring that a node is not directly connected to another node.
5. Declaring that two nodes are independent, given a condition-set.
6. Providing partial nodes ordering, that is, declare that a node appears earlier than another node in the ordering.
7. Providing a complete node ordering.

D. Decision Tree Algorithm

Decision tree induction algorithms, an inductive learning task use particular facts to make more generalized conclusions. Most decision tree induction algorithms are based on a greedy top-down recursive partitioning strategy for tree growth. They use different variants of impurity measures, like; information gain [22], gain ratio [23], and distance-based measures [24] to select an input attribute to be associated with an internal node. One major drawback of
Greedy search is that it usually leads to sub-optimal solutions. A predictive model based on a branching series of
Boolean tests, these smaller Boolean tests are less complex than a one-stage classifier. Entropy of decision tree is the information gain measure, is minimized when all values of the target attribute are the same, If we know that commute time will always be short, then entropy = 0. Entropy is maximized when there is an equal chance of all values for the target attribute (the result is random),
If commute time = short in 3 instances, medium in 3 instances and long in 3 instances, entropy is maximized.
Calculation of entropy:

$$\boldsymbol{S} = \sum_{i=1}^{l} \frac{|-\boldsymbol{S_i}|}{|\boldsymbol{S}|} \log_2 \frac{|\boldsymbol{S_i}|}{|\boldsymbol{S}|}$$

S = set of examples
Si = subset of S with value vi under the target attribute
l = size of the range of the target attribute.

Decision tree classifier is able to break down a complex decision making process into collection of simpler and easy decision. The complex decision is subdivided into simpler decision. It divides whole training set into smaller subsets. Information gain, gain ratio, gain index are three basic splitting criteria to select attribute as a splitting point. Decision trees can be built from historical data they are often used for explanatory analysis as well as a form of supervision learning. The algorithm is design in such a way that it works on all the data that is available and as perfect as possible [25]. There are many specific decision tree algorithms such as ID3 (Iterative Dichotomiser 3), C4.5 Algorithm, Successor of ID3, CART (Classification And Regression Tree), MARS: extend decision tree to better handle numerical data.

V. CONCLUSION

The selection of efficient feature and its reduction from the huge dataset is a tedious task and it is not necessary that all features are informative. So it becomes very essential in many application areas such as remote sensing, image retrieval, intrusion detection etc. In this paper three feature selection methods and some classification algorithm is discussed. In these methods and algorithm some are relevant to feature selection and reduction without redundancy. It is concluded that the feature selection methods are helpful in the dimensionality reduction from the huge database due to this performance of the learning affected and it is not necessary that all methods and algorithm is suitable for feature selection and reduction some has drawbacks. So in future work implement such algorithm for feature selection which is most appropriate for eliminating the redundancy and irrelevant feature.

REFERENCE

[1]. Hoai Bach, Bing Xue, Ivy Liu and Mengjie Zhang, "Filter based Backward Elimination in Wrapper based PSO for Feature Selection in Classification", *IEEE Congress on Evolutionary Computation (CEC)*, pp. 3111-3118, 2014.

[2]. J.C.Bezdek, Pattern Recognition with fuzzy objective function algorithms. Plenum Press, New York, 1981.

[3]. J.Kang, L.Min, Q.Luan, "Novel modified fuzzy c-means algorithm with applications", *Digital Signal Processing*, 2009, **19**: 309-319.

[4]. K.Dunne, Cunningham and F.Azuaje, "Solution to instability problems with sequential wrapper-based approaches to feature selection", *Journal Of Machine Learning Research*,2002.

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

[6]. J.G. Dy, and C.E. Brodley, "Feature selection for unsupervised learning," *J. Mach.Learn. Res*. **5** pp. 845–889, 2004.

[7]. H. Liu, and J. Ye, "On similarity preserving feature selection", *IEEE Trans. Knowledge Data Engg*. Vol. **25**, pp. 619–632, 2013.

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

[9]. Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Department of Computer Science & Engineering, Arizone State University, Tempe, AZ 85287-5406, USA, 2003

[10]. A.Blum and P.Langley, "Selection of relevant features and examples in machine learning", *Artificial Intelligence*, vol **97**, pp 245-271, 1997.

[11]. H. Liu and L. Yu. "Toward integrating feature selection algorithms for classification and clustering",. *IEEE Transactions on Knowledge and Data Engineering*, **17**(4):491, 2005.

[12]. S. Ma and J. Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, **9**(5):392–403, 2008.

[13]. Y. Saeys, I. Inza, and P. Larranaga. "A review of feature selection techniques in bioinformatics" *Bioinformatics*, **23**(19): 2507–2517, 2007.

[14]. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. "Gene selection for cancer classification using support vector machines", *Machine learning*, **46**(1-3):389–422, 2002.

[15]. J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[16]. J. R. Quinlan "Induction of decision trees" *Machine learning*, **1**(1):81–106, 1986.

[17]. RAJ, M. A., Bincy G, Mrs. T. Mathu. "Survey on common data mining classification Technique", International Journal of Wisdom Based Computing.

[18]. Aci, M., Inan, C., and Aveit, M., "A hybrid classification method of k nearest neighbor, Bayesian methods and genetic algorithm, *Expert Systems with Applications"*, Vol.**37**, pp.5061-5067, 2010.

[19]. Yildiz, T., Yildirim, S., and Altilar, D.T., "Spam filtering with parallelized KNN algorithm", Akademik Bilisim, 2008.

[20]. Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", *Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009*, March 18 - 20, volume I, Hong Kong.

[21]. M. Soundarya, R. Balakrishnan, "Survey on Classification Techniques in Data mining", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. **3**, Issue 7, July 2014.

[22]. Barros, R. C., Basgalupp, M. P., De Carvalho, A. & Freitas, A. A. "A survey of evolutionary algorithms for decision-tree induction". Systems, Man, and Cybernetics, Part C: Applications and Reviews, *IEEE Transactions on*, **42**, 291-312.

[23]. Wang, Y., Makedon, F. S., Ford, J. C. & Pearlman, J, "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data". *Bioinformatics*, **21**, in 2005 1530-1537.

[24]. De Mántaras, R. L. "A distance-based attribute selection measure for decision tree induction" *Machine learning*,1991, **6**, 81-92.

[25]. Pawar, T., Kamalapur, S. "A Survey on Privacy Preserving Decision Tree Classifier".