CHAPTER 4

FRAMEWORK OF DATA REDUCTION METHOD

This chapter deals with the comparison framework of different preprocessing filters implementation on different clustering and classification methods to identify better data reduction group to represent the very large data sets using various techniques like clustering, classification and selection of attributes in data mining. It deals with the detailed implementation made towards the research goal. The system development involves the comparison framework to identify the performance of various preprocessing filters on a large data set. The data set then treated with various data mining techniques like clustering, classification and selection of attributes to identify the robustness (i.e.) the technique finds the better data reduction grouping when compared with other technique. This comparison and implementation of the algorithm have been done in five consecutive phases. The chapter deals with the basic implementation of the different preprocessing filters on the large volume dataset, performance comparison of different clustering methods using selection of attributes, comparison of different classification methods application, combining clustering and classification techniques and identifying and locating the best data reduction groups in the large data set which helps in minimizing the large volume of data set effectively and efficiently.
4.1 ARCHITECTURE OF PROPOSED DATA REDUCTION METHOD

The framework of the Data Reduction method has various preprocessing filters and grouping techniques, as depicted in Figure 4.1. This architecture is based on an honest model of finding a better preprocessing filter with grouping technique among the various available filters and grouping techniques that identifies the best features of the entire large volume of data set.

The large volume of dataset is initially treated with preprocessing filters like Normalize, Randomize from different levels like instance and attribute. Now the dataset is free from missing data, incorrect data and redundant data. The improved dataset is clustered using different clustering techniques, classified using different classification techniques to find the better data representatives of the large volume of dataset which is the proposed work.

![Figure 4.1 The Framework of the proposed data reduction method](image)
4.2 PREPROCESSING FILTERS

This research work mainly focuses on the preprocessing filters which preprocess the dataset and make the dataset suitable for grouping that aims to achieve feasible data reduction. Normally any generated report and query result from the database tool very often result in large and complex dataset. The preprocessing of data helps in filtration of missing, irrelevant and repetitive data which does not affect the performance of the dataset. Redundant or impartial pieces of data can confuse or disorient a user. Preprocessing the data using filters can make the data more efficient and suitable for data reduction. There are different types of preprocessing filters are available that can be used to alter the reports, query results or any other kind of information results.

4.2.1 Filters

The filter method always precedes (Bangsuk Jantawan & Cheng-Fa Tsai 2014) the clustering / classification technique. The output of the filter approach is used as the input to different clustering and classification algorithms. The filter method (Sudha & Jebamalar Selvi 2017) selects the subset of the data on the fundamental characteristics which is not dependent on the data mining algorithm.

4.2.1.1 Filters of Clustering

4.2.1.1.1 Normalize

The Weka (3.7.4) tool consists of preprocessing technique which consists of set of filters in both supervised and unsupervised learning. The Normalize filter is the one which present in the attribute level. The Normalize filter normalizes all the numeric values in the selected dataset. The resulting values must be in [0,1] by default for the given dataset which is used to calculate the normalization intervals in Randomize filter.
4.2.1.1.2 Normalize

The Weka (3.7.4) tool consists of preprocessing technique which consists of set of filters in both supervised and unsupervised learning. The Randomize filter is the one which present in the instance level. The randomize filter randomly shuffles the order of data instances passed through it. The random number generator resets the seed value whenever the new set of dataset is passed in Normalize filter.

4.2.1.1.3 Normalize

The Weka (3.7.3) tool is the initial version that consists of preprocessing filters both supervised and unsupervised learning. In this version the Normalize filter presents in both the attribute and instance level. But in the later version it gets modified. The Normalize filter normalizes all the numeric values in the selected dataset. The resulting values must be in [0,1] by default for the given dataset which is used to calculate the normalization intervals.

4.2.1.2 Filters of Classification AddClassification

A Filter named addclassification (Bangsuk Jantawan & Cheng-Fa Tsai 2014) for adding the classification and the class distribution and an error flag to a dataset with a classifier. The classifier is either trained on the data itself or provided as a successive model. The proposed classifier methods are compared in the following section.

4.3 CLUSTERING

Data mining plays a vital role in terms of prediction and analysis (Grabmeier & Rudolph 2002). In data mining the clustering is a kind of technique which is used to group the data on the basis of similarity. These groups are often called as Clusters. The clustering algorithms are useful in the fields like data mining, pattern recognition, learning theory (Priyanka Sharma
2015) to identify the clusters in the given dataset. Clustering is an unsupervised learning technique used for grouping elements, the elements in the same group are more similar than the elements in the other groups. In (Mor M Gupta & Sharma 2014) exploratory data mining the clustering is the main task, in statistical data analysis the clustering is the common method and used in many fields like machine learning, pattern recognition, image analysis, information retrieval, world wide web and bioinformatics. Cluster analysis was originated in anthropology by (Driver and Kroeber 1932) and introduced to psychology by (Zubin 1938) & (Robert Tryon 1939). The datasets with large number of features and size, makes a learning model to over fit and therefore learning performance degenerates.

Clustering is the assignment of grouping a set of objects in a group and therefore the objects in the same group are more similar to each other the objects in other groups. Clustering basically implements mutual information based iterative procedure like vector quantization thus may be sensitive to initial setting. Clustering is the center of attention of active research in a number of fields such as statistics, pattern recognition and machine learning. The idea of data reduction is to find data subset by completely removing the features with little or no predictive information. Therefore the data reduction helps in improving the performance of any learning models by minimizing the dimensions to any desired number, providing with the generalization capability and also increases the speed of the learning process. It also helps in acquiring the better understanding of the datasets by showing the most important features and the relationships exist between them.

Data reduction using clustering in unsupervised learning is a challenging task due to the presence of irrelevant attributes and high dimensions. This paper shows the performance of the comparison of different clustering algorithms based on the application of filters and usage of selection of attributes and finally comparing the results identifies which performs better
The K-Means algorithm is the one which performs the best when compared with other clustering algorithms. The different clustering algorithms works differently for different data sets. WEKA (Waikato Environment for Knowledge Analysis) is the one surrounded by the various data mining tools which clusters the datasets in unsupervised learning. Clustering is an unsupervised learning which is used to decide the intrinsic groups in a set of unlabeled data. Grouping of objects is prepared based on the standard of maximizing the intra-class relationship and minimizing the inter-class relationship in such a way that the objects in the same cluster share some similar properties. In general, they can be classified as wrapper and filter methods according to the evaluation condition in searching for appropriate features (Bangsuk Jantawan & Cheng-Fa Tsai 2014). For wrapper approaches the excellence of every candidate data subset is assessed by investigating the performance of an exact clustering algorithm on this subset, and each candidate subset is obtained by conducting combinatorial explore through the space of all data subsets. These algorithms have shown the success on low dimensional data. The Filter approach is described in the instance and attribute level shows the better performance by filtering irrelevant, missing incomplete and redundant data in the given data set. It can be considered as either an investigative task or preprocessing step. If the goal is to preprocess and explore the hidden patterns in the data, clustering becomes a standalone preprocessing exploratory task by itself.

In machine learning, the data reduction using clustering can be viewed as the mixture of various explore techniques for proposing novel data subsets, along with an valuation measure which scores differently for different data subsets. Data reduction using clustering is used for three reasons:
• simplification of models to make them easier to interpret by researchers/users
• short executing training and testing times
• enhanced generalization by reducing over fitting

Data reduction using clustering should be distinguished from data extraction. Data extraction creates new features from the original features, whereas data reduction returns a subset of the features. The data reduction using clustering along with special preprocessing tool called Filters initially remove all the features that are redundant, irrelevant, incomplete and missing value without incurring much loss of information. When clustering is carried out on such datasets, the results prove the greater performance. Redundant and irrelevant features are two distinct notions, since one relevant data may be redundant in the occurrence of another relevant data with which it is strongly correlated. Before applying the clustering, the preprocessing method called filters are applied on the dataset in the removal of redundant and irrelevant features with not much loss of information.

Clustering along with data selection aims to choose a small subset of the relevant features from the original ones according to certain relevance evaluation criterion which usually leads to better learning performance. A data reduction algorithm can be seen as the combination of a search technique for proposing new data subsets, along with an evaluation measure which scores the different data subsets. The unsupervised data selection becomes more difficult in the absence of class labels that can guide the search for the relevant information. Clustering is an unsupervised learning problem which is used to determine the intrinsic grouping in a set of unlabeled data. Grouping of objects is done on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity in such a way that the objects in the same cluster share some similar properties or traits. They can be
classified as wrapper and filter methods according to the evaluation criterion in searching for relevant data. For wrapper approaches the quality of every candidate data subset is assessed by investigating the performance of a specific clustering algorithm on this subset, and each candidate subset is obtained by conducting combinatorial search through the space of all data subsets. These algorithms have shown the success on low dimensional data. It can be seen as either an exploratory task or preprocessing step. If the goal is to preprocess and explore the hidden patterns in the data, clustering becomes a standalone preprocessing exploratory task by itself.

4.3.1 Expectation Maximization Algorithm

An Expectation Maximization algorithm EM enables parameter estimation in probabilistic models with incomplete data. An EM algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. An EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step (Dampster et al. 1977). These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.
4.3.2 K-Means Algorithm

K-Means clustering is basically a partitioning method used to explore data and treats explanation of the data items based on locations and distance linking various input data points. Partitioning the items into mutually private clusters (K) is done by it in such a fashion that items within each cluster stay as close as possible to each other but as far as possible from items in other clusters. Each cluster is identified by its centre point i.e. centroid. The distances used in clustering in the majority of the times do not really mean the spatial distances. In general, the only conclusion to the problem is the discovery of universal minimum which is the exhaustive choice of starting points. But use of a number of replicates with arbitrary starting point leads to a solution i.e. a universal solution. In a dataset, a preferred number of clusters K and a set of K initial points and the K-Means clustering algorithm identify the favorite number of distinctive clusters and their centroids.

A centroid is the point whose co-ordinates are obtained by means of calculating the average of each of the co-ordinates of the points of samples assigned to the clusters.

Algorithmic steps for K-Means clustering

1) Set K – To decide a number of preferred clusters, K.

2) Initialization – To decide k initial points which are used as first estimates of the cluster centroids. They are taken as the initial first values.

3) Classification – To inspect each point in the dataset and allocate it to the cluster whose centroid is adjacent to it.

4) Centroid calculation – When each point in the data set is
assigned to a cluster, it is again needed to calculate the new K centroids.

5) Convergence process – The steps of (iii) and (iv) needs to be carried out until no point changes its cluster assignment or until the centroids no longer moves.

The exact data samples should be collected before the application of the clustering algorithm. Preference has to be given to the features that explain each data sample in the database. As in the other clustering algorithms, K-Means requires that a distance metric between points is to be definite. This distance metric is used as a measure in the above mentioned step (iii) of the algorithm. A common distance metric used is the Euclidean distance. In case, when the diverse features datasets are used in the data vector which have different relative values and ranges, the distance computation may be unclear and therefore it has to be scaled. The input parameters for the clustering algorithm are the number of clusters that are to be established along with the initial starting point values. When the initial values are given, the distance from each sample data point to each initial value is calculated using equation. Then each data point is placed in the cluster linked with the nearest starting point. Finally, when all the data points are assigned to the cluster, then the new cluster centroids are calculated. For each factor in the cluster, the new centroid is calculated. The new centroids are then considered as the new initial starting point values and steps (iii) and (iv) of the algorithm are repeated again. This process continues until any more data point changes or until the centroids no longer move.

4.3.2.1 Advantages of K-Means Clustering

In large data sets most of the time the K-Means algorithm is computationally faster than other cluster algorithms, when the value of k is
chosen to be small. The K-Means produce tighter clusters than other existing clustering algorithms, especially when the cluster are globular.

4.4 SELECTION OF ATTRIBUTES

The selection of attributes is technique that selects the set of attributes which is sufficient to represent the entire dataset. This technique finely picks the attributes that acts as the representatives for the remaining attributes which are not selected. This selection of attributes consists of two processing methods like attribute evaluator and the search method. The attribute evaluator consists of the collection of evaluator methods to suits to the type of dataset being used. Based on the attribute selector the secondly there are variety of search methods available that are also implemented on the selected dataset based on its type.

- Attribute Evaluator
- CfsSubsetEval

This evaluator measures the implication of a subset of attributes by taking into account the individual prognostic ability of each data beside with the degree of redundancy between them.

The subsets of features that are highly associated with the class, while having low inter connection are preferred.

- Search Method
- Best-First

Best-first search is a search algorithm which explores a graph by extending the most competent node selected according to the specified rule.
"Best-first search" refers knowingly to a search with a heuristic that attempts to forecast how close the solution to the end of a path, so that paths which are judged to be closer to a solution are fixed first. This specific type of search is called greedy best-first search or pure heuristic search to a specified rule.

- Greedy Step Wise

A greedy algorithm is an algorithmic method that follows the method of heuristic of constructing the most nearby favorable choice at each stage with the trust of finding a global optimum. Generally, in many problems, a greedy policy does not produce an optimal solution, but however a greedy heuristic may yield local optimal solution that reasonably find an accurate global optimal solution in a rational time.

4.5 CLASSIFICATION ALGORITHM

In Supervised Classification the set of possible class labels known in advance. Classification performs data reduction in the form of forecasting a certain outcome based on a given input using predefined class labels (Daiho Uhm, et al. 2012). In order to predict the outcome, the algorithm process a training set containing a set of features and the respective outcome, called as goal or predicted data. The algorithm tries to discover relationships between the features that would make it possible to forecast the outcome.

Classification – A Two-Step Process

4.5.1 Model Construction

Describes a set of pre determined classes. Each sample is assumed to belong to a predefined class as determined by the class label data. The set of features used for the model construction is the training set. The model is the representation of the any type of classifier and its method used for
classifier construction.

4.5.2 Model Usage

The model usage classifies the future or unknown objects and estimates the correctness of the model. The known data label of the test sample is compared with the classified result of the training model obtained from the model. Accuracy rate is the percentage of the test set sample that are correctly classified by the model. The test set is independent of the training set. If the accuracy is acceptable, then use the model to classify the data features whose class label is not known. Therefore, in supervised learning the training data are accompanied by labels indicating the class of observations and the new dataset is classified based on the training set.

4.5.3 Definition of Classifier

The classifier is used for classification. When using the classifier, the data features used to estimate the accuracy of the classifier. The classification rules can be applied to the new data features if the accuracy considered acceptable.

4.5.3.1 Factors of Classifier

i. Accuracy – Accuracy of classifier refers to the ability of classifier in forecasting the possible features which makes the large dataset. It forecasts the class label of the features correctly and the accuracy of the classifier refers to how well a given classifier can guess the value of forecasted feature for a new data.

ii. Speed – This refers to the computational cost in generating and using the classifier or predictor.
iii. **Robustness** – It refers to the ability of classifier or predictor to make correct forecasts from the given noisy data.

iv. **Scalability** – Scalability refers to the ability to construct the classifier or predictor efficiently to handle the given large volume of data.

v. **Interpretability** – It refers to what extent the classifier or predictor understands about the given dataset and generate forecasts based on the user requirement.

4.5.4 **Meta Classifier**

A meta-classifier is a classifier that doesn’t implement any classification algorithm on its own, but uses another classifier to do the actual work. In addition, the meta-classifier adds another processing step that is performed before the actual base-classifier sees the data.

4.5.5 **Classifier and its Types**

In classification algorithm there exists the various classifier algorithms. The meta classifier and tree classifier are the existing classifier which performs better than other classifiers. The Meta classifier is one among the various classifier methods of the classification algorithm. This meta classifier consists of two techniques like Bagging and Random subspace.

4.5.5.1 **Bagging**

Bootstrap aggregating is otherwise called as bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms which is used for classification. It reduces variance and improves the predictive force, therefore helps to avoid over fitting. Bagging is a special kind of the model averaging approach. Given a standard training set D of size n, bagging generates m new training sets Di,
each of size \( n' \), by sampling from \( D \) uniformly and with replacement. By sampling with replacement, some observations may be repeated in each \( D_i \). If \( n' = n \), then for large \( n \) the set \( D_i \) is expected to have the fraction \( (1 - 1/e) \approx 63.2\% \) of the unique examples of \( D \), the rest being duplicates. This kind of sample is known as a bootstrap sample. The \( m \) models are fitted using the above \( m \) bootstrap samples and combined by averaging the output for regression or voting for classification. Bagging leads to improvements for unstable procedures like artificial neural networks in classification.

4.5.5.2 Random Subspace

In machine learning the random subspace method is also called as attribute or feature bagging, is an ensemble learning method that attempts to reduce the correlation between estimators in an ensemble by training them on random samples of features instead of the entire data set. In ensemble learning, one tries to combine the models produced by several learners into an ensemble that performs better than the original learners. The random subspace method is similar to bagging except that the features or attributes are randomly sampled, with replacement, for each learner.

4.5.6 Tree Classifier

The tree classifier is one among the various classifier methods of the classification algorithm. This tree classifier consists of two techniques like Random Forest and REPTree.

4.5.6.1 Random Forest

Random forest normally refers to randomly built trees which have nothing to do with machine learning. Random forest is a popular machine learning framework where the tree constructed with nodes at random.
data mining tool WEKA uses the Random forest to refer to a assessment tree built on a random subset of columns.

4.5.6.2 REPTree

REPTree uses the regression tree logic and creates several trees in different iterations. After that it selects greatest one from all generated trees. That will be considered as the representative. In pruning the tree, the measure used is the mean square error on the predictions made by the tree. REPTree is a fast decision tree learner which constructs a decision tree using information gain as the splitting principle, and prunes it using reduced error pruning. It only sorts values for numeric features once. Experimental results are presented in section 5 and the concluding remarks forms the final section of the paper.

4.6 PROPOSED WORK - IMPLEMENTATION METHODOLOGY

For the purpose of comparing and analyzing the efficiency in terms of total time taken in seconds (secs) to build the model with and without using filters namely 1. Normalize with respect to attribute type and 2. Normalize with respect to Instance type in Unsupervised filter. In preprocessing the filter named “Normalize” normalizes all numeric values in the given dataset. Then the corresponding dataset with normalized numeric values are executed with the algorithms namely EM and K-Means using WEKA, the well-known PSD Machine Learning Repository is used for the comparative study. It is actually a collection of databases which is widely used by the researchers of Machine Learning. Livestock Dataset: Total number of attributes is twelve of which all are numeric. The total numbers of instances are 106331 Poultry Dataset: Total number of attributes is twelve of which all are numeric. The total numbers of instances are 30924.
Similarly comparing and analyzing the efficiency of data reduction of a large dataset in terms of total time taken in seconds (secs) to build the model using before and after application of selection of attributes on a dataset that is preprocess using filters, on which two clustering algorithms are executed then the results are recorded and compared, identifies a better data reduction of the given large dataset. The two clustering algorithms are executed with application of two filters namely 1. Normalize with respect to attribute type and 2. Randomize with respect to Instance type in unsupervised learning. The preprocessing the filter named “Normalize” normalizes all numeric values in the given dataset and randomize “Randomly shuffles the order of instances passes through it.

During the process the corresponding dataset is preprocessed using filters and executed with Expectation Maximization clustering algorithm and K-Means clustering algorithm which has undergone the before and after application of selection of attributes technique. The given dataset of numeric values is executed with the algorithms namely EM and K-Means in WEKA (Priyanka Sharma 2015), the well known PSD Machine Learning Repository is used for the comparative study.

The large dataset is treated with two classifiers like Meta classifier and Tree classifier are computed on a livestock dataset to analyze their performance in terms of data reduction (the minimum time taken to construct the model) when making use of the classifier method like bagging and random subspace in Meta classifier and RANDOM FOREST and REPTree in Tree classifier to identify the competent classifier for classification. For the purpose of comparing and analyzing the efficiency in terms of data reduction by identifying the class labels in a big dataset the two classifiers in Classification algorithm is used. The data are observed in the table format are in the type of i) computing initially classifier method on the dataset ii) firstly
applying the filter called add classification method on the dataset and then using classifier methods to identify the classes for the dataset which enables to identify the usage of the filter. The two classifier methods like Meta and Tree are used on the dataset. During the computation the corresponding dataset is executed with two classifiers without and with the appliance of filters. The given dataset of numeric values are executed with the classifier algorithms in WEKA, the well known PSD machine learning repository is used for the comparative study. It is really a set of databases which is extensively used by the researchers of Machine Learning. Livestock Dataset: Total number of attributes is twelve of which all are numeric. The total numbers of instances are 106331.

The Data mining algorithms classification and clustering perform arrangement of data into groups based on their unique style of processing Here in this paper a csv format livestock dataset is processed classification and clustering algorithms of data mining. The dataset is initially treated with filters and then with both the algorithms and finally the performance is represented in terms of minimum time taken to build the model in secs is identified. Here in this paper the Meta classifier with RSS method add classification filter generates the very minimum time to build the model when compared with the other classification algorithms.