CHAPTER 2

RELATED WORKS

The previous works are highlighted by showing the links to the results obtained in this study. The main objective in this work is grounded on data preprocessing. Despite more than two decades of continuous development learning from variety of data is still a focus of deep research. With the expansion of algorithms in machine learning combined with the arrival of big data era, there emerge new challenges. Recent trends are towards analyzing not only the distribution of classes, but also other difficulties in the inherent nature of data. Researchers Bartosz Krawczyk, (2015) discuss these open issues and challenges that need to be addressed to further develop the field of imbalanced data distribution. Seven vital areas of research in this topic are identified, covering the full spectrum of learning from imbalanced data: classification, regression, clustering, data streams, big data analytics and applications, e.g., in social media and computer vision.

This chapter discusses the foundations of this research study in the form of related articles and it is organized as follows. This chapter is organized as follows. Section 2.1 presents data size issues and data preprocessing. Section 2.2 presents various works seen in machine-learning in health care. Section 2.3 provides the evidences data preprocessing by over-sampling and cost sensitive techniques. Section 2.4 adds applications in the context of Software Engineering, Internet and Mobile. Section 2.5 enumerates the objectives of this study. Section 2.6 declares the contributions made finally.
2.1 Data size issues and Data preprocessing

Data Preprocessing happens to be one of the most important issues within the well-known Knowledge Discovery from Data process. Data directly taken from the source will likely have inconsistencies, errors or most importantly, it is not ready to be considered for a data mining process. These topics are covered by the authors García et al. (2015) and furthermore, the increasing amount of data in recent science, industry and business applications, are shown. Data preprocessing techniques are scalable even for ‘Big Data’, a term which is used to describe massive amount in the form of documents, emails and web pages. The pre-processing algorithm (Ashish R. et al. 2014) to extract real time user accessed data and to mine and analyze leads to deeper insights into business patterns and trends. Another area for the demand for data preprocessing is web mining to clean the data and user identification process to identify unique users. Here RanjenaSriramand Mallika.R (2016) an efficient pre-processing technique and an innovative Hashing technique to identify Distinct User for Web Usage Mining.

Data preparation tasks are refined and enhanced to the industrial standards by recent authors Joaquin Perez et al. (2015). A method known as ‘Cross-Industry Standard Process for Data Mining (CRISP-DM)’ is meant for such tasks. Data preparation can be of General and Specific types for epidemiological domain. The work by Vu Anh Tran et al. (2014), noise and outliers are removed by corresponding attraction and density in the data and it is called D-IMPACT data preprocessing which is an iterative technique yielding better results in two dimensional datasets.

Feature selection is performed in various ways either by supervised or unsupervised technique often include redundant or correlated features. This yields extra cost and removing redundancy completely can make the system vulnerable to
measurement errors by Pan et al. 2008, and Mishra et al. (2009). Other researchers, like Banerjee Monami (2015) propose a new scheme for feature selection in unsupervised category labeled as ‘UFeSCoR’ for avoiding the non-significant features and allowing the significant ones. Hence this approach helps to control the redundancy comparing around five other feature selection methods. A novel feature weighting scheme ‘GRW’ is provided (SagarImambi.S 2011, Amirasayed et al 2013) and proved to support the classification accuracy while testing on medical data set.

2.2 Machine-learning in health care

Among all the application domains, the most benefited and hence enhanced its insight is the health care. It depends on huge diagnosis and clinical data for its knowledge discovery for extracting the patients’ patterns for reliable detection systems at earlier stage and other various healthcare related systems. Evolution of such exercises are enumerated by Neesha Jothia et al. (2015). Other areas across many walks of life, like manufacturing, education, financial modeling, policing, and marketing are adopted with data-intensive machine-learning methods arisen from AI studies prevailing in commerce, science and technology, aiming at stronger evidential decision-making. This fact is thrown with more light by Jordan et al. (2015). In the field of predicting pressure ulcer, the high risk factors during long surgery happen to be serious problem. The learning models are obtained using the familiar four techniques in data mining. They are typically Mahalanobis Taguchi System (MTS) by Mahalanobis distance metric, Support Vector Machines (SVMs) by margin distance to highlight support vectors, decision trees by predicates and leaf to yield decision and finally by simple line fitting by logistic regression are applied to do the prediction of invasion of pressure ulcers. This has been verified and successfully obtained by Su, C. T et al. (2012). As an alternative to methodologies
already on offer the data mining technologies are suggested for different disciples meant for social science methodologies, such as cognitive science, psychology and human behavior. Enormous growth of such technologies are studied and reported by Liao Shu-Hsien et al. (2012). The data collected pertaining to five chronic diseases: hypertension, diabetes, cardiovascular disease, disease of the liver, and renal disease are applied to establish early-warning criteria that may reduce the complication occurrence rate.

Nested and ten folded cross validation are included in experiments performed and, following the results obtained, the optimal results are shown by the K-Nearest Neighbors algorithm: more than 81% of accuracy and 78% of AUC (Area Under the Curve) Receiver Operator Characteristic curve, which happens to be a very good result in this scenario. Lei Tang et al (2009). The different tumors were collected using the development of clinical technologies and features have been collected for breast cancer diagnosis.

The hybrid approach given by Zheng et al. (2014) using K-means and Support Vector Machine (one may call K-SVM) is suggested in this context. Most of the time, filtering all the significant features to support the clinical disease diagnosis is challenging and time consuming task. One such task was performed at the University of California over data set from Wisconsin Diagnostic Breast Cancer (WDBC) - Irvine machine learning repository.

As a result six tumor features out of 32 existing features are searched from the training set and applied to improve the accuracy up to 97.38%, by Zheng et al. (2014). The familiarity of application of SVM is reiterated again while treating the people with type 2 diabetes is mostly based on managing drug therapies, aiming at monitoring glucose levels properly.
2.3 Over-sampling and cost sensitive techniques

K-nearest neighbor algorithm, Linear Discriminant method, and forward sequential selection are utilized to get 80% accurate predicting model for detecting such Chronic disease which has steadily become a major death cause in Taiwan as made studies by Jen, C et al. (2012) In another context simple models logistic regression (LR) and decision tree (DT) are established by merging synthetic minority over-sampling technique (SMOTE) ,cost-sensitive classifier technique (CSC), bagging, boosting and under-sampling along the attribute subset identification is used to select relevant attributes to detect survivability prediction of breast cancer. This work is reported by Wang, K et al. (2013).

Another instance of studies on Breast cancer by Garcia-Laencina et al. (2015) is also to detect the survival of breast cancer patients. The most disadvantage state is having the absence of information, i.e., missing data, in certain clinical trials. However, most standard learning techniques are not able to cope up incomplete data and, then, missing data imputation is a generally applied strategy for solving this inconvenience.

Therefore, and taking into account the characteristics of each breast cancer dataset, it is required to analyze the environment with four types of imputations by first without any imputation, by mode, by expectation maximization (EM) and KNN imputation. Prediction models for survivability of breast cancer are made use by four different methods: Logistic Regression, K-Nearest Neighbors, Decision Trees and Support Vector Machines.
2.4 Context of Software Engineering, Internet and Mobile

Iterative and multiple partitioning based filter are tried for data preprocessing in the context of Software Engineering. This has been improving software quality prediction by Taghi M et al. (2007). The final results achieved with models built on the datasets based on filtered fitting are more accurate than the models built on the dataset with higher noisy level.

Experimental results by Seokho Kang et al. (2015) proposed a novel method that takes less time for training to achieve reasonable success, relative all the traditional SVM ensembles. The penetration of Data mining techniques for prediction in Intrusion Detection system for providing Protection was analyzed by Jayamagarajothi et al. (2015) for presenting the dramatic increase in the internet applications, security which is becoming a major issue of the network.

The contribution of multi objective evolutionary algorithms used for data mining is indicated by Mukhopadhyay et al. (2014). This adds more light at the dimension towards feature selection. While the data size becomes voluminous 'crowd sensing' harnesses the power of the crowd by mobilizing a huge groups of mobile users attached with various networked devices to gather data with the basic multi-modal and large-volume features. Various data mining techniques are applied for modeling by Zhang et al. (2014).

2.5 Objectives

Proposed framework contains four components in order to tackle the performance issues through four aspects namely Data size, Attribute Selection, Cost Ratio and Cross Validation applied for medical data sets like diabetes and cancer clinical environments. The objectives for presenting these aspects are as follows:
➢ To demonstrate the significance of data size (both column and row sizes) in the model performance generated by possible different tools.

➢ To reduce the complexity due to ‘curse of dimensionality’ by attribute selection are made with subset evaluation methods and searching procedures like best first search, greedy search, etc.

➢ To establish the role of cost conscious instead of being blind at learning and to obtain minimum error or maximum accuracy

➢ To induce the most accurate model and to estimate the true model's accuracy on future (unseen) instances the cross validation procedures are applied with more diversity and randomness. the generalization aspect of the learning with different

2.6 Contributions

This research study in the background of machine learning or data mining is focusing on data preprocessing issues for conducting experiments and obtaining results in the domains especially critical areas in medical field like diabetes, cancer, etc.

First key for safe start of data mining is data preprocessing and it has many aspects researched for long time. In this study contributions are made towards particularly on the ‘DACC’ aspects. Final results are achieved in order to throw more light on these aspects carrying the influence of data size (D), selection of the significant attributes (A), following different strategies of cross validation (C) and the inclusion of the cost ratio(C) are the four aspects considered this thesis. The techniques like meta classification, modified cross validation, cost sensitive learning and extended attribute selection are applied in order to verify the hypotheses in the
proposed framework based on DACC aspects. DACC aspects are shown in the proposed architecture diagram (Fig 2.1).

This research work is based on diabetes dataset and the training and testing data sets are downloaded from American Diabetes Association where much research around diabetes prediction has been conducted. The main objectives are to predict what variables are the causes, at high risk, for diabetes and to provide a preventive action toward individual at increased risk for the disease. Many attributes have been reported in literature as important indicators for diabetes prediction. However the accuracy for recommendation for assisting the physician is challenging. The clinical variables such as BMI, blood pressure, glycaemia, cholesterol, or cardio-vascular risk are included for constructing the result.

The era of data science is accelerated with its storage and computational capacities. The performance issues are still under research as novel methods and
schemes are evolving. The framework in this research is focused around data preprocessing stage in constructing learning models for predictions in the domain of diagnosing diseases like diabetes. The overview of the contents in this research work in simple descriptions without giving lengthy details. The first issue considered is the size of the data set. In general it determines the performance of the learning models as depicted in the following Fig 2.2.

![Fig. 2.2 Accuracy Vs Training Set Size](image)

In this study, data size issues are tried in two directions. Firstly small data sets with smaller number of columns and secondly small number of instances (records). The context of diseases in fruits is considered. The main intention in this context is to deal with available ‘small’ data set for high accuracy. For this purpose AdaBoost, Bagging, Logit Boost models are built using an open source mining Weka under supervised learning algorithms. It is necessary to reduce the error before constructing the final models and thus the varying the parameters and number of iterations for training is carried out.

Moreover the efforts for enlarging the number of columns by class membership functions and artificial attributes give better results. Data quantity is the
main issue in the small data set problem, because usually insufficient data will not lead to a robust classification performance. Extracting more effective information from a small data set is thus of considerable interest. Here there is a new attribute construction approach which converts the original data attributes into a higher dimensional feature space to extract more attribute information by a similarity-based algorithm using the classification-oriented fuzzy membership function. Seven data sets with different attribute sizes are employed to examine the performance of the proposed method. The results show that the proposed method has a superior classification performance when compared to principal component analysis (PCA), kernel principal component analysis (KPCA), and kernel independent component analysis (KICA) with a Gaussian kernel in the support vector machine (SVM) classifier.

The next important aspect of data pre-processing is attribute selection as the number of attributes plays an important role in constructing the learning model. Search methods are applied to select the subset of attributes. The dataset for this context is used from diabetes diagnosis. Search methods like Best first search, Greedy Search are applied to get the final list of attributed before feed into training process. After selecting the attributes classification accuracies are obtained and compared. The research study is further augmented by not only the attribute selection but also the parameter training and proper selection of Kernels.

The third aspect in this framework is novel as rare research is so far available in the studies of relationship between the cost ratio of false positives to that of false negatives to the accuracy of learning models.

Cost sensitive models as well the cost blind models are usually both accepted by their performance through minimum error or maximum accuracy. Hence in
critical context of diagnosing diseases like diabetes and cancer need to be applied with more meaningful measures such as ρ. These findings can support the decisions better for physician diagnosing with more confidence.

The cross validation involves iteration over number of folds, i.e. parts of training data and testing data for getting the model and the accuracy. It has been proposed the method of cross validation through the randomising the selection of train and test data at each iteration step.