International Journal for Rapid Research in Engineering Technology & Applied Science
ISSN (Online): 2455-4723
Vol 2 Issue 7 August 2016

Data Quality Assessment & Enhancement for classifier's Performance Improvement

Shubhangi sharma[1],Maya yadav[2]

[1]M.Tech Research Scholar Sanghvi Institute of Management and Science. Indore, India

[2]Assistant Professor Sanghvi Institute of Management and Science. Indore, India

[1]shubhangisharma91@gmail.com , [2]maya.yadav@sims-indore.com

Abstract:

The data mining is a tool which includes various mathematical definitions and methodologies by which the data is corrected, manipulated, and identified according to their patterns. Additionally that is frequently used for various data classification, pattern recognition and other decision making tasks. The learning of data mining techniques are basically depends on the datasets or the training samples. According to the training samples the learners can be classified as supervised and unsupervised models. Additionally when the supervised learning is used the data sets are played more essential roles to guide the learning algorithm. But due to incomplete, random and inconsistent data can affect the performance of supervised classifiers. Therefore the proposed study work is concentrated over improving the data set quality using the pre-analysis of data. In order to find most optimum data analysis algorithm various techniques and methodologies are studied. But most of them have limitations such as limited class processing ability and a number of data processing steps. Thus a new data is proposed in this study using the linear regression analysis for analysing the data. The proposed algorithm works in three major steps first input data noise estimation using regression analysis, second noisy data separation and the finally the data correction. In these steps the data is analysed and unfit patterns or noisy patterns those are affecting the performance is detected and corrected using the outlier estimation technique. The implementation of the proposed algorithm is provided using the MATLAB environment. After implementation the performance of proposed and available technique is evaluated and compared over different datasets of machine learning. The comparative performance of the system demonstrates effectiveness of the proposed algorithm. According to the obtained results the proposed algorithm improves the classification accuracy between 5-25% additionally reduces the resource consumption approximately 5%. Therefore the proposed technique is adoptable due to high accuracy and less time consumption..

Keywords— **performance improvement, data mining, data quality, classification, implementation**

I. INTRODUCTION

Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Data mining is used to extract information from a dataset and transform it into an understandable structure for further algorithmic use. The real world applications such as financial data analysis, retail industry, biological data analysis, telecommunication industry and many more are utilizing the techniques of automated data analysis. According to the applications and the environment of learning the data mining methods may be categorized as either supervised or unsupervised. In unsupervised methods, no target variable (class labels) identified. The data mining algorithm responsible to search patterns among the entire feature sets. The unsupervised data mining method is also termed as clustering algorithms.

Supervised methods of learning refer to a function that learns the data patterns from labelled training data. Training data includes both the input and the desired output in terms of attributes and their class labels. The common supervised data mining methods are classification, regression etc. but the predefined patterns for the learning, is not much accurate or that contains the noisy in patterns, then the learner is not properly trained for the desired task. On the other hand if the data is well defined and accurately provides the patterns for the required outcomes then the

learning algorithm properly works and learns most of the patterns produced for training. Therefore the proposed work is intended to presents a data evaluation approach for improving the quality of data to optimizing the classification patterns of classifiers. The goal of proposed system is to modify the learning set in such manner by which the classifier accurately predicts the target classes.

Data mining is a domain of information and knowledge processing where the raw data is analysed and processed for recovering the informative patterns form the data. According to the available literature the data mining process are working in three main phases the data pre-processing, data model development and finally the performance analysis and testing. In first phase the available raw data is manipulated for cleaning and refining the data for optimizing and correcting the available information. In second steps using the pre-processed data algorithm works and produces the machine learning data model. Finally the data model is tested and performance is evaluated through different real world scenarios.

According to the data processing and machine learning steps, in each phase of data mining the data is modified for enhancing the patterns hidden in data. Therefore after pre-processing some of the researchers are implementing the feature selection and feature extraction approaches. The low level features are responsible to enlarge the patterns and their significance. The proposed work is intended to enhance the data mining based learning process by filtering the noisy patterns in training samples. Basically the data is found inconsistent, random and noisy in nature therefore the raw data is not suitable for learning. Thus such kind of noisy data not provides the appropriate significance for target information (class labels). Thus to improve the training data significance the feature extraction and noise reduction is a primary goal of the proposed study. The main reason behind the concept is the data quality is affecting the performance of classifiers performance. Incomplete data and/or noisy data are not suitable for accurate data modeling. Low quality data produces the incorrect learning that results poor classification capability of machine learning algorithm.

II. PROPOSED TECHNIQUE

The machine learning and accurate data mining technique involves a number of issues and challenges. The key issue is to identify the meaningful features to train the learning algorithms by which accurate classification or pattern recognition becomes feasible. This proposed study is focused to improve the dataset quality and training patterns for enhancing the classification accuracy of the available classifiers. Therefore the following issues are targeted to find the solution.

- In machine learning processes, it is hard to make accurate classification because of less number of features in data sets or contains the incomplete information about the attributes.
- Data quantity and quality is the main issue in the datasets if the attributes of the patterns are distributed randomly then the accurate predictions are not much feasible
- A datasets may be containing irrelevant, redundant information instances that are known as noise or the outliers. These noises or irrelevant information introduces difficulty in learning processes. That affects the classification accuracy after training or data modeling.

Suppose a data set contains the instances of data such that

$$D = \{I_1, I_2 \dots I_n\}$$

Where each object in this set of data having a set of attributes $I_i = \{a_1, a_2, \dots a_n\}$. These attribute having a relationship with each other to define a target pattern or a class $C = \{C_1, C_2 \dots C_m\}$. Therefore a classification algorithm is made just like a function to identify the relationship among the attributes to prepare the data model. This trained or learned data model accepts a set of symbols (i.e. participating attributes $a_1, a_2, \dots a_n$) and using the prepared model classify the set of attributes in some predefined classes $C = \{C_1, C_2 \dots C_m\}$ such that

$$model.classify(a_1, a_2, \dots a_n) \rightarrow C$$

If all the attributes having a relationship with corresponding class values then a target class is predictable, but when the amount of noise in attribute set is higher or the amount of attributes are less than a suitable definition of class values are not properly defined. For instance in place of $a_1, a_2, \dots a_n$ the input is produces $a_1, b_2, \dots a_n$ (b_2 is noise element) or $a_1, a_2, \dots a_{n-1}$ (n-1 shows the missing values) then the predicted class labels are not properly recognized. Thus a function is additionally required to measure the noise and undefined relationships to improve the quality of data for appropriate learning.

In order to find the solution for the above defined complexity the following work is included in the proposed system.

- If an accurate classification required thus is need to increase the features or significance of attribute relationships over the datasets. For this purpose feature construction or data refinement is required.

- Pre-processing or missing values handling or noise estimation or feature enhancement or feature selection of data is required.

Proposed system includes simple steps for data quality assessment and improvements data quality. Therefore a simple process is demonstrated using figure 1. The given figure includes flow diagram and steps to understand the simplicity and effectiveness of the proposed technique.

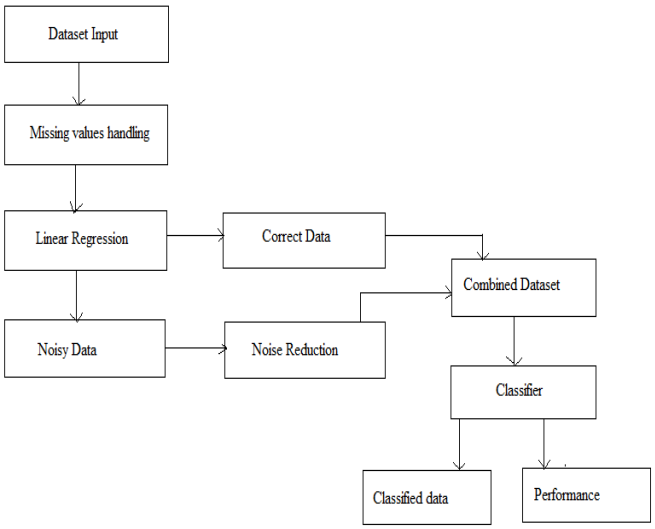

Figure 1 proposed system

Dataset input

The training data is produced in this phase as input to the system. The dataset can be structured or unstructured in formats but the in this approach the structured dataset is considered. This data set is contains the set of attributes and a pre-defined class labels for making it use for the supervised learning algorithms. For experimentation purpose the WEKA supported datasets (i.e. iris, wine,..) are used.

Missing values handling

The dataset used are recognized as the 2D vectors, these vectors are having N number of instances (rows) and M number of attributes (columns) these attributes are grouped according to the pre-defined patterns or class labels. But sometimes the set of attributes contains the null values or missing values thus refinement on the dataset is required. The procedural steps of handling the missing attributes are summarized on the below given table 1.

Input: dataset D

Output: refined dataset R_d

Process:

1. [col, row] = readDataset(D)
2. $for\ (i = 1; i \leq row; i++)$
   a. $for\ (j = 1; i \leq col; j++)$
      i. $if(D[i][j] == null)$
         1. remove $D[i]$
      ii. end if
   b. End for
3. End for

Table 1 missing values handling

Linear regression

Regression analysis is a statistical method of finding relationship between the data attributes. In linear analysis the data is modeled using linear approximation function. The key advantage of working with the regression analysis is that, it allows additional inputs and outputs relevant to statistical analysis. The outcomes of the linear regression is a least-squares estimator, lower confidence bounds, upper confidence bounds, residuals, matrix of intervals, and the statistics (i.e. the R2 statistic, the F statistic and p value, and error variance).

Basically in simple linear regression analysis, initially dataset is a set of n points, in this context an independent variable x_i, and two parameters, $\beta_0\ and\ \beta_1$ are used for class predations thus the relationship is developed linearly. And these n points are going to fit in a straight line therefore the:

Straight line can be defined using $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n$……. (1)

Adding a term in x_i^2 to the preceding regression gives:

Parabola: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad i = 1, \dots, n$…… (2)

This is still linear regression; although the expression on the right hand side is quadratic in the independent variable x_i, it is linear in the parameters $\beta_0, \beta_1\ and\ \beta_2$. In both cases, ε_i is an error term and the subscript i indexes a particular observation. Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$y'_i = \beta_0 + \beta_1 x_i \text{…….. (3)}$$

The residual, $e_i = y_i - y'_i$ is the difference between the value of the dependent variable predicted by the model y'_i, and the true value of the dependent variable y_i. One method of estimation is ordinary least squares. This method obtains parameter estimates that minimize the sum of squared residuals.

Therefore the regression analysis is made using the function:

$$[\beta_0, \beta_1, R, R_i, stat] = regress(D) \ldots\ldots\ldots\ldots\ldots\ldots (4)$$

Where β_0 is used for least-squares estimator, confidence intervals are defined by β_1. R is stands for residuals and matrix of intervals is given by R_i. $stat$ is used for statistics.

Noisy data:

The previous step's outcome two parameters $R, and\ R_i$ are used for to estimate the outliers from the input data. That can be performed using rcoplot function in MATLAB. Thus a function is used

$$recoplot(R, R_i) \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

The figure 3.2 is generated using the equation 5. Because R_i is a two dimensional matrix which contains the matrix of intervals and R shows the residuals. In this graphical representation red line shows the higher error than defined confidence interval. Therefore that is an error among the available data or an outlier. Thus that is required to find out the index of the input dataset. by which the outliers are removed from the data.

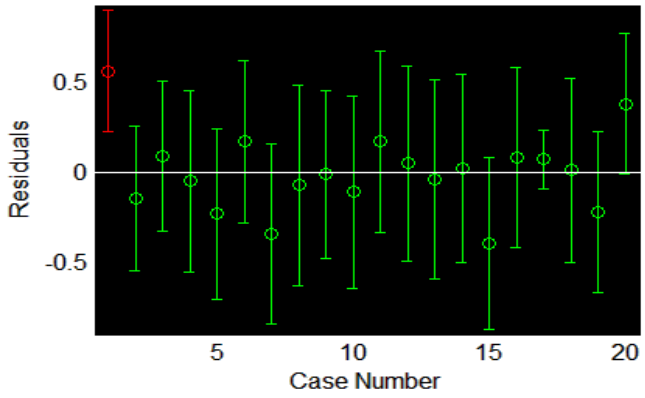

Figure 2 outlier detection

Noise reduction

The identified outliers from the data are removed in this phase thus the light weight process for the length of the dataset size is prepared. That accepts the R_i two dimensional matrix and evaluation of data is performed using the following functional algorithm.

Input: matrix of intervals R_i, The data set D Output: A cleaned and corrected dataset D_c
Process: 1. $for\ i = 1\ to\ D.length()$ 2. If $R_{i,1} \leq 0\ and\ R_{i,2} \leq 0$ 3. $D_c = \frac{D(a_i)+\frac{1}{N}\sum_{i=1}^{N} D(a_i)}{2}$ 4. Else if $R_{i,1} \geq 0\ and\ R_{i,2} \geq 0$ 5. $D_c = \frac{D(a_i)+\frac{1}{N}\sum_{i=1}^{N} D(a_i)}{2}$ 6. End if 7. End for 8. Return D_c

Table 2 data correction

Correct data

The corrected data D_c is the final outcome of the proposed methodology for enhancing the features of the data to improve the classifier's performance. Now the experimental environment for classification with the supervised learning is prepared. In this approach the SVM classifier is used.

Classifier

In machine learning, SVM is supervised learning models with associated learning algorithms. SVMs belong to a family of generalized linear classifiers and can be interpreted as an extension of the perceptron. SVMs are a group of supervised learning methods that can be applied to classification or regression. It is primarily a two class classifier. SVMs can efficiently perform non-linear classification using what is called the kernel function; indirectly map their inputs into high-dimensional feature spaces. It can also solve multiclass problem with the help of kernel methods and kernel function. It aims to maximise the width of the margin between classes, that is, the vacant area between the decision boundary and the nearest training pattern. The basic idea of SVM classifier is to choose the hyper plane that has maximum margin. The dashed lines drawn parallel to the separating line mark the distance between the dividing line and the closest vectors to the line. The space between the dashed lines is called the margin. The vectors (points) that constrain the width of the margin are the support vector. Suppose the

two classes can be presented by two hyper planes parallel to the optimal hyper plane.

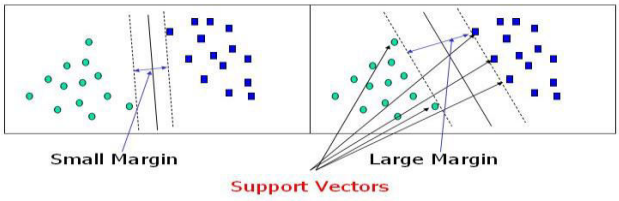

Figure 3 Support Vectors

$$w_n x_t + b \geq 1 \; for \; y_t = 1 \; and \; t = 1,2,3 \ldots, k$$

$$w_n x_t + b \leq -1 \; for \; y_t = -1$$

Where w = {w_1, w_2, w_3,....., w_n } is a vector of n elements

Figure 3 represents the small margin, large margin and support vectors during classification of a two class dataset. The kernel method consists of two modules: First one is the choice of kernel and the second one is the algorithm which takes kernel as input. The basic idea of kernel method is to map the data from input space to feature space F using Ø [11], Ø: X → F where X = "inputs", F = "feature space", Ø = "feature map". The space of the original data is called input space We say that k(x, y) is a kernel function if there is a feature map Ø such that for all x, y K(x, y) = Ø(x) . Ø (y). In pattern recognition a feature space is an abstract space where each pattern sample is represented as a point in n-dimensional space. Its dimension is resolute by the number of features used to describe the patterns. The concept of a kernel mapping function is very powerful. It allows SVM models to perform partings even with very complex boundaries. Figure 3.4 shows that how to map the data from low dimensional space to higher dimensional space.

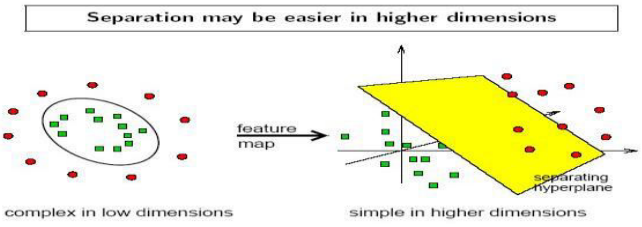

Figure 4 Non-linear Separable Data

The mapping function needs to be computed because of a tool called kernel trick. The kernel trick is a mathematical tool which can be applied to any algorithm which solely depends on the dot product among two vectors. Every place a dot product is used; it is substituted by a kernel function. When appropriately applied, those candidate linear algorithms are transformed into non-linear algorithms. Those non-linear algorithms are correspondent to their linear originals operating in the range space of a feature space Ø. However, because kernels are used, the Ø function does not need to be ever explicitly computed. Kernel functions must be continuous, symmetric, and most rather should have a positive (semi-) definite Gram matrix. Kernels which are said to satisfy the Mercer's theorem are positive semi-definite, meaning their kernel matrices has no non-negative Eigen values. A positive definite kernel insures that the optimization problem will be convex and solution will be unique. Types of kernel functions:

1. **Linear Kernel:** The Linear kernel is the simplest kernel function. It is given by the inner product <x, y> plus an optional constant **c**.

$$k(x, y) = (x^T y + c)$$

2. **Polynomial Kernel:** The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well appropriate for problems where all the training data is normalized.

$$k(x, y) = (\alpha x^T y + c)^d$$

Adaptable parameters are the slope **alpha**, the constant term **c** and the polynomial degree **d**.

3. **Gaussian kernel:** The Gaussian kernel is an example of radial basis function kernel.

$$k(x, y) = e^{\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)}$$

Classified data

In this phase of data processing the data is classified using data mining algorithms i.e. SVM.

Performance

In this phase the performance of the system is evaluated and reported using the different performance parameters i.e. accuracy, error rate, time consumption and space complexity over the original dataset as that are available and processed according to the proposed technique.

III. RESULTS ANALYSIS

The performance of the similar classification algorithms over two different data correction approaches are evaluated and compared in this chapter. Therefor the different performance factors are evaluated and provided i.e. accuracy, error rate, memory consumption and time complexity.

A. Experimental Datasets

In data mining and machine learning the key element of data models are data. Therefore according to the data utilization for learning is also affected the classification and categorization. There are two kinds of machine learning datasets are available first those dataset includes the class labels with the attribute sets these data sets are termed as supervised datasets. And secondly the dataset not including the target classes are known as unsupervised datasets.

The most frequently used datasets for machine learning datasets are available at the machine learning dataset repository in arff format. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. ARFF files have two distinct sections. The first section is the Header information, which is followed the Data information. The Header of the ARFF file contains the name of the relation, a list of the attributes (the columns in the data), and their types.

B. Accuracy

The performance of the classifier in terms of accuracy is given in this section. The performance evaluation of classifier is evaluated using n cross validation technique. The accuracy of the system can be given using the following formula.

$$accuracy = \frac{total\ correctly\ classified\ samples}{total\ samples\ avilable} X100$$

Dataset	Simple dataset	Traditional method	Proposed method
Pima dataset	64.8438	77.8646	84.82
Australian dataset	91.5942	94.4928	95.2
Bupa dataset	66.3768	71.0145	78.8
Fourclass dataset	81.2065	89.3271	94.7
Iris dataset	95.3333	98.6668	99.62
Seed dataset	93.3333	93.3333	95.2

Table 3 classifier accuracy

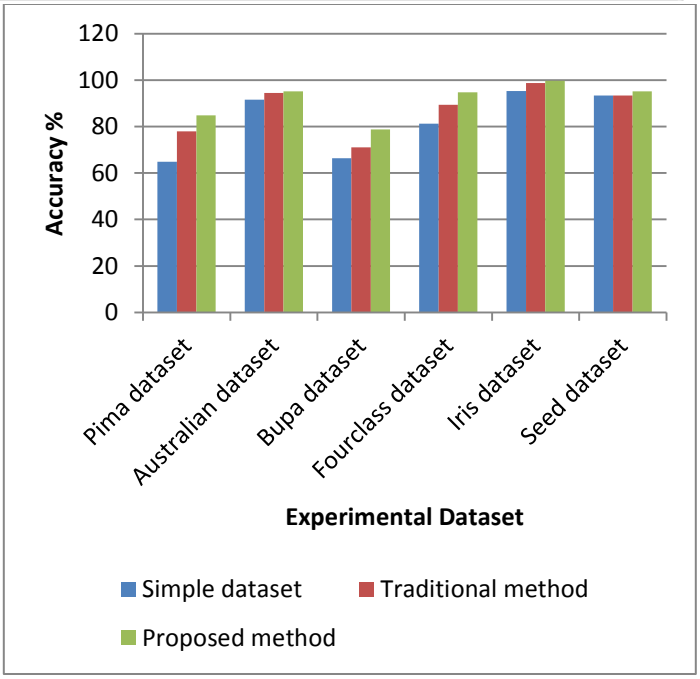

Figure 5 Classifier accuracy

The performance of the proposed system and traditional approach is provided using the figure 5. In all the experimentation the similar classifier namely SVM is used but the datasets are changed according to the method selected. The blue bars in the given figure contain the SVM performance for the original datasets without any change. The red line of the figure demonstrates the performance of the fuzzy based data evaluation and feature construction technique. Finally the green line shows the performance of the data corrected by the proposed approach. According to the obtained results the classifier perform much better learning for proposed data quality enhancement technique as compared to the traditional technique.

C. Error rate

The error rate of the classifier reports the amount of data that are not properly recognized during the classification. The error rate of the classifiers can be evaluated using the following formula.

$$error\ rate = \frac{misclassified\ patterns}{total\ patterns\ to\ classify} X100$$

Or

$$error\ rate = 100 - accuracy$$

Figure 6 and table 4 shows the error rate of the implemented systems. Therefore to represent the

performance of the algorithms the X axis contains the different experimental datasets and the Y axis shows the error rate percentage. The performance of the simple or original data set is given using blue bars and the red bar shows the performance of the traditional fuzzy based approach finally the green bars shows the performance of the proposed technique. According to the obtained performance the proposed technique produces the less error rate as compared to the traditional approach therefore the proposed technique is much effective as compared to the traditional approach.

Dataset	Simple dataset	Traditional method	Proposed method
Pima dataset	35.16	22.14	15.18
Australian dataset	8.41	5.51	4.8
Bupa dataset	33.63	28.99	21.2
Fourclass dataset	18.8	10.68	5.3
Iris dataset	4.67	1.333	0.38
Seed dataset	6.67	6.67	4.8

Table 4 error rate

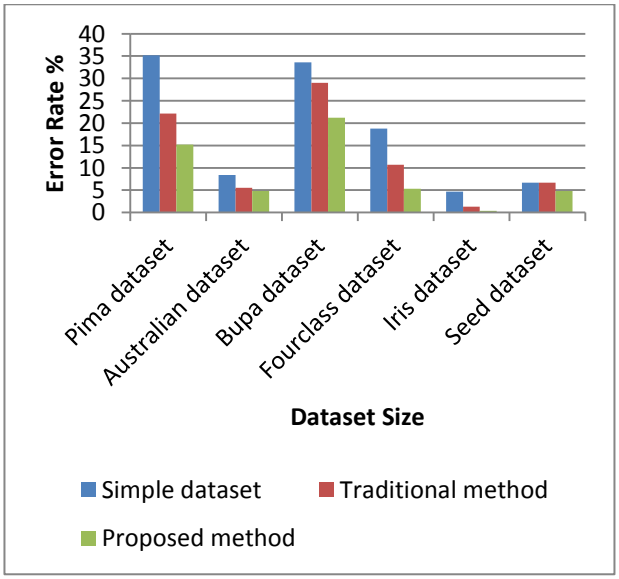

Figure 6 error rate

D. Memory Consumption

Memory consumption is also termed as space complexity of the system. The amount of main memory consumed during the system execution is known as the memory consumption or space complexity. The obtained memory consumption during different experiments is given using table 5 and figure 7. The performance of the experimental scenarios using the SVM classifier and the original dataset is given using the blue line. Similarly the performance of fuzzy based technique is demonstrated using the red line and finally the proposed technique is simulated using the green line according to the obtained performance the proposed technique consumes less memory for correcting the dataset as compared to the traditional approach.

Dataset	Simple dataset	Traditional method	Proposed method
Pima dataset	32171	35193	34216
Australian dataset	30291	35281	34251
Bupa dataset	32811	35928	34332
Fourclass dataset	33921	34726	30271
Iris dataset	30548	32947	31261
Seed dataset	34927	35872	34266

Table 5 memory consumption

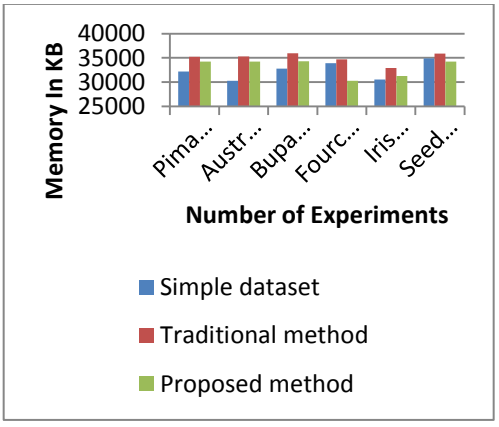

Figure 7 memory consumption

E. Time Consumption

The amount of time is required to generate the outcomes from the algorithm is known as the time complexity or time consumption of the system. The obtained results with the experimentation is given using table 6 and figure 8

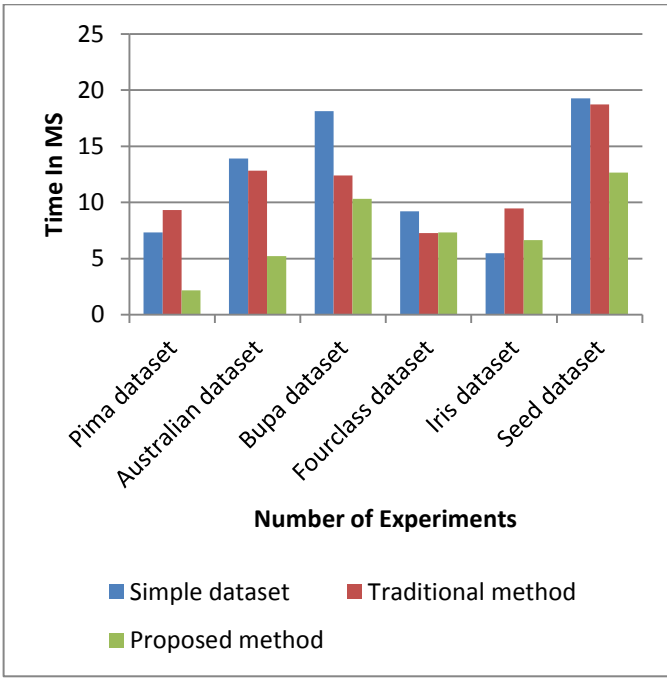

Figure 8 Time consumption

Number of experiments	Simple dataset	Traditional method	Proposed method
Pima dataset	7.31	9.33	2.16
Australian dataset	13.91	12.81	5.21
Bupa dataset	18.11	12.38	10.32
Fourclass dataset	9.21	7.26	7.31
Iris dataset	5.48	9.47	6.63
Seed dataset	19.27	18.72	12.66

Table 6 Time consumption

The obtained performance of the proposed and traditional algorithm for dataset quality improvement technique is demonstrated using the figure 8 and table 6 in this diagram the X axis shows the data set used during the experiments and the Y axis shows the time complexity of the algorithms for correcting the dataset attributes. According to the obtained performance the proposed algorithm consumes less amount of time as compared to the traditional approaches of the data correction.

IV. CONCLUSIONS

The entire work on data quality improvement and feature correction is performed successfully the chapter provides the essential facts obtained during the observations and experiments. Additionally the future extension of the proposed technique of data analysis is also reported in this chapter.

A. Conclusion

Data mining offers a way by which the hidden patterns form the raw input data are discovered. This process need to develop some algorithm that analyse the data and formulate them in a mathematical model. The formulated model helps to recognize the similar pattern data. But the noisy data can affect the formulation of data model and the decision making capability of the developed classifier model. Proposed study addresses the problem of noisy and incomplete datasets and their performance issues. Therefore a new technique is developing which pre-evaluate the data and their patterns for obtaining the incompleteness and noisy candidates of the datasets.

Thus the proposed work includes the missing or null values handling technique by evaluating data instances. Additionally for approximating the composition of noise in the datasets a linear regression based technique is presented. The linear regression technique helps to find the outliers from the data using the difference intervals and the estimated residuals. The estimated noisy instances of the data are corrected in the next phase. For correction of the noisy or outlier instances the mean algorithm is used and the quality of data is improved. The mean method enable the data to lies among the normal distributed points.

The implementation of the proposed data quality assessment and improvement technique is performed using the MATLAB environment. After implementation the performance of the proposed system is evaluated and compared with the traditional method of performance enhancement of classifier by enhancing the datasets. For the comparative performance study of the system the accuracy, error rate, time consumption and the memory used. The obtained performances of the similar classifier by enhancing the quality of data sets using both the techniques a performance summary table 7.

S. No.	Parameters	Traditional method	Proposed technique
1	Accuracy	Low	High
2	Memory	High	Low
3	Time complexity	High	Low
4	Error rate	High	Low

Table 7 performance summary

For evaluating the performance of the algorithms a number of different datasets are used and their quality improvement is performed. After enhancing the quality of dataset the similar classifier is used for classification and the performance is obtained. The obtained performance shows the proposed technique effectively reduces the misclassification rate and improves the accuracy of the classifiers.

B. Future work

The proposed system is adoptable and efficient according to the demonstrated performance analysis over different datasets and similar classifier. The proposed method is suitable for data quality estimation and data feature enhancement for supervised learning approaches. In near future the proposed technique can be enhanced for the following domains.

1. Feature estimation and dimensionality reduction techniques
2. Outlier detection in unsupervised learning algorithms i.e. clustering techniques.

REFERENCES

Der-Chiang Li and Chiao-Wen Liu, "Extending Feature Information for Small Data Set Classification," IEEE Transaction on Knowledge and Data Engineering, vol. 24, No. 3, March 2012.

Akash Shrivastava, Kuntal Barua, "An Efficient Tree based algorithm for Association Rule Mining", International Journal of Computer Applications (0975 – 8887) Volume 117 – No.11, May 2015

Data mining Concepts and Techniques, Second Edition, Jiawei Han and Micheline Kamber, http://akademik.maltepe.edu.tr/~kadirerdem/772s_Data.Mining.Concepts.and. Technique-s. 2nd.Ed.pdf

A Comparison of Several Approaches to Missing Attribute Values in Data Mining, Jerzy W. Grzymala-Busse and Ming Hu, Springer-Verlag Berlin Heidelberg 2001, pp. 378−385, [5] Data Mining - Applications & Trends, http://www.tutorialspoint.com/data_mining/dm_applications_trends.htm

Mahak Chowdhary, Shrutika Suri and Mansi Bhutani, "Comparative Study of Intrusion Detection System", 2014, IJCSE All Rights Reserved, Volume-2, Issue-4

Mrs. Pradnya Muley, Dr. Anniruddha Joshi, "Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence", International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163, Issue 4, Volume 2 (April 2015)

Ghazaleh Khodabandelou, Charlotte Hug, Rebecca Deneckere, Camille Salinesi, "Supervised vs. Unsupervised Learning for Intentional Process Model Discovery", Business Process Modeling, Development, and Support (BPMDS), Jun 2014, Thessalonique, Greece. pp.1-15, 2014

Ritika, "Research on Data Mining Classification", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014

Abbas Jafari, S. S. Patile, "Use Of Data Mining Technique to Design a Driver Assistance System", Proceedings of 7th IRF International Conference, 27th April-2014, Pune, India, ISBN: 978-93-84209-09-4

Fabricio Voznika, Leonar Doviana, "Data Mining Classification", http://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf

M.S.B. PhridviRaj, C.V. GuruRao, "Data mining – past, present and future – a typical survey on data streams", The 7th International Conference Interdisciplinarity in Engineering (INTER-ENG 2013), 2013 The Authors. Published by Elsevier Ltd

Luca Belmonte, Rosanna Spera and Claudio Nicolini, "SpADS: An R Script for Mass Spectrometry Data Pre-processing before Data Mining", J Comput Sci Syst Biol, Volume 6(5)298-304 (2013) – 298

Massimiliano de Leoni, Fabrizio M. Maggi, Wil M.P. van der Aalst, "An alignment-based framework to check the conformance of declarative process models and to pre-process event-log data", & 2013 Elsevier Ltd All rights reserved.

Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Pre-processing in Web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore

Swasti Singhal, Monika Jena, "A Study on WEKA Tool for Data Pre-processing, Classification and

Clustering", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013

Victoria López, Alberto Fernández, Jose G. Moreno-Torres, Francisco Herrera, "Analysis of pre-processing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics", Expert Systems with Applications, 2011 Elsevier Ltd. All rights reserved.

Salvador Garcı´a, Julia´n Luengo, Jose´ Antonio Sa´ez, Victoria Lo´pez, and Francisco Herrera, "A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 4, APRIL 2013

Alberto Fernández, Victoria López, Mikel Galar, María José del Jesus, Francisco Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches", Knowledge-Based Systems, 2013 Elsevier B.V. All rights reserved.

Jose-Norberto Mazón, Jose Jacobo Zubcoff, Irene Garrigós, Roberto Espinosa, Rolando Rodríguez, "Open Business Intelligence: on the importance of data quality awareness in user-friendly data mining", EDBT-ICDT '12 Proceedings of the 2012 Joint EDBT/ICDT Workshops Pages 144-147 ACM New York, NY, USA ©2012

Hongzhou Sha, Tingwen Liu, Peng Qin, Yong Sun, Qingyun Liu, "EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining", Information Technology and Quantitative Management , ITQM 2013