Kurdistan Journal of Applied Research (KJAR) | Print-ISSN: 2411-7684 – Electronic-ISSN: 2411-7706 | kjar.spu.edu.iq
Volume 2 | Issue 3 | August 2017 | DOI: 10.24017/science.2017.3.2

Comparative Study of Classification Techniques For Large Scale Data - Case Study

Nigar M. Shafiq Surameery
Building and Construction Engineering Dept.
College of Engineering
University of Garmian
Kalar, sulaimani, Iraq
nigar.mahmoud@garmian.edu.krd

Dana Lattef Hussein
Database Dept.
Computer Science Institute
Sulaimani Polytechnic University
Sulaimani, Iraq
dana.hussein@spu.edu.iq

Abstract: ***The existence of Massive datasets that are generated in many applications provides various opportunities and challenges. Especially, scalable mining of such large-scale datasets is a challenging issue that attracted some recent research. In the present study, the main focus is to analyse the classification techniques using WEKA machine learning workbench. Moreover, a large-scale dataset was used. This dataset comes from the protein structure prediction field. It has already been partitioned into training and test sets using the ten-fold cross-validation methodology. In this experiment, nine different methods have been tested. As a result, it became obvious that it is not applicable to test more than one classifier from the (tree) family in the same experiment. On the other hand, using (NaiveBayes) Classifier with the default properties of the attribute selection filter has a great time consuming. Finally, varying the parameters of the attribute selections should be prioritized for more accurate results.***

Keywords: classification techniques, WEKA, data mining, bioinformatics, knowledge discovery, large-scale data.

1. INTRODUCTION

Nowadays, a large amount of data is being gathered and processed. The traditional way is manually accumulating the data but this task becomes uninteresting in the case of huge amounts of data (Pavlidis, et al. 2002). Computers have brought about substantial improvements to technology that helps to deal with enormous amounts of data (Bergmann, Jan and Naama 2003) (Fayyad and Paul 1997). The new technologies make it possible to use and organize the huge volumes of data (Angus-Hill, et al. 2001) (Kifaya 2009). Managing large-scale data has become a major field of research, namely data mining (Li and Wong 2002).

Data Mining is a process of identifying the unique and the required patterns in large - scale data. Its learning techniques can be classified in to both supervised and unsupervised. A common unsupervised technique is clustering (Huttenhower, et al. 2006). While the common supervised learning techniques, which are useful to be used in medical and clinical research, are Classification, Association rules and Statistical regression (Schreiber and Baumann 2007).

This study will focus on the usage of different classification techniques on large-scale bioinformatics dataset. The WEKA software has been used to perform the classification process. This is because it consists of the major learning techniques that can be used in classifying and analysing massive amounts of medical data such as Bayesian classifiers and decision trees. (Yang, et al. 2007).

2. LITERATURE REVIEW

In this part, the most relevant works to this study have been reviewed. The concentration of the review has been given to the aim of the work, the classifier types and the results.

The study conducted by (AL-Nabi and Ahmed 2013), comparing the performance efficiency among three classification's techniques (Decision tree, KNN, Bayesian) and analysing the required time complexity using a Survey. As a result, the authors argued that all DecisionTree's algorithms are the easiest algorithm as compared to KNN and Bayesian algorithms and have the least error rate. The result showed that the DecisionTree outperformance and Bayesian classification had the same accuracy while the other methods that based on clustering, such as KNN Classification, are not giving good results. The study supposed that while using KNN classifier, it is possible to improve the efficiency of results by increasing the number of datasets while increasing the attributes can affect the efficiency of the result for Bayesian algorithm classifier.

Another group of researchers (David, Saeb and Al Rubeaan 2013) have conducted a study for analysing Data Mining Tools and Classification algorithms in Medical Bioinformatics. The medical bioinformatics analyses have been performed to demonstrate the usage of WEKA in the diagnosis of Leukemia. "leukemia_all_72x7129" database, which has 7130 attributes and 72 instances, has been used for comparing the (Decision tree J4.8, Bayesian Network and a Naïve Bayes) classification techniques depending on their reported accuracy, learning time and the error rate. During the experiment, the highest accuracy reported was 98.6111% for Bayesian in 0.17 seconds and the lowest was 81.9444% for Decision tree.J48 in 0.62 seconds.

Another research done by (Tan, et al. 2010) was concentrated on analysing huge volume data mining and compares their performance to the MapReduce model. The authors used the Naive Bayes classification technique with different methods dealing with the training set. Using a real-world huge volume data set in their research, they explored and compared the accuracy of three large-scale data mining approaches. The result of the experiments shows; for building a more accurate model, more data in training should be used. In other words, with increasing the sampling size, the sampling model can also get the same accuracy as the ones built on the available data. Moreover, the authors explored the relationship between the accuracy of the model with the sampling size in sampling model and the model accuracy with the data partitioning size in ensemble model. Based on their analysis, they proposed an idea of utilizing MapReduce framework to help improve the accuracy and efficiency of sampling.

3. METHODS AND MATERIALS

3.1. WEKA SOFTWARE

In this study, WEKA software has been represented as useful tool for performing the classification mechanism, pre-processing and analysis of result (Fayyad and Paul 1997) (Erica and Falk 2009). It represents a collection of machine learning algorithms written in Java programming language; provide the required tools to obtain a better understanding that allows the biological problems to be solved. The Name of WEKA software stands for "Waikato Environment for Knowledge Analysis". It is available under the General Public License. "Weka software was developed at the University of Waikato in New Zealand". (Schreiber and Baumann 2007).

As shown in figure (1), the Weka software can be started from one of the four different interfaces on the main GUI Chooser window. These interfaces are Explorer, Experimenter, Knowledge Flow and Simple CLI. The following is a simple explanation for each of them (Tan, et al. 2010) (Kretschmann, Fleischmann and Apweiler 2001).

Simple CLI: "simple command-line" interface that can be used for the direct execution of WEKA commands on operating systems that do not provide their own command line interface (Pavlidis, et al. 2002).

Knowledge Flow: which represents is an interface that creates a flow of information by connecting the individual learning components graphically (Kifaya 2009).

Experimenter: This is an "environment for conducting statistical tests between learning schemes and performing experiments that could be distributed across more than one computer running remote experiment servers". (Yoo, et al. 2012)

Explorer: This is the main interface in Weka; through the explorer, users can load data in various formats ARFF, C4.5, CSV, and library (Li and Wong 2002). In the present study, the explorer interface was used to load the data and applying the classification process.

As shown in figure (2), there are six (6) tabs in the WEKA explorer window, which can be used to perform different tasks such as pre-process, classify, associate, cluster, select attribute and visualize.

3.2. CLASSIFICATION TECHNIQUES

Classification techniques can be used to process and analyse a huge volume dataset. It takes each instance of the data and assigns it to a specific class in order to decrease the classification errors. This procedure is used as recognized method for making the same decisions in any new situations (Guerra, et al. 2011) (Eisen, et al. 1998).

The classification process applied in two-steps. In the first step, The Classification algorithm applied on training data set to create a model. While in the second step; the extracted model is tested against a predefined test dataset in order to measure the accuracy and performance of the model. (John and Langley 1995) (Luscombe, Greenbaum and Gerstein 2001). As result, the Classification techniques can be defined as, "the process to assign class label from such a dataset whose class label is unknown". (David, Saeb and Al Rubeaan 2013).

3.3. BRIEF DESCRIPTION OF THE DATA SET

In this experiment, there is a large-scale bioinformatics dataset that comes from the protein structure prediction field. It has been partitioned into training and test sets using the ten-fold cross-validation methodology adapted to protein datasets. Each training set has about half 234000 instances. The dataset has 180 attributes and there is no missing value to be treated in this experiment. The train set details (instances, class distribution, attributes and type of attributes) has been provided in Table 1, while Table 2 contains the same details for the test set.

From table (1) and table (2), it is clear that there are (180) attributes for each fold of the data set, and all of them are Numeric. Regarding to the class distributions, each fold has only 2 types of classes with different amplitude (number of instances for each class).

3.4. EXPERIMENTAL DESIGN

In this section the combinations of methods, which have been used to design the experiment and solve the problem will be explained. By using [Weka] software, the experiment has been easily designed and (9) different methods have been tried. In all of the (9th) methods the following components have been used:

-The (ArffLoder) from the (data sources tab) that is used to add the data sets and loading the data.

-The (classAssigner) from the (evaluation) tab which is used to specify which column to be the class.

- The (TrainingSetMaker) from the (evaluation) tab, that is used to make a data set into a training set.

-The (TestSetMker) from the (evaluation) tab, that is used to make a data set into a test set.

-The (ClassifierPerformanceEvaluator) from the (Evaluation) tab, which is use to evaluate the performance of patch trained/tested classifier.

-The (TextViewer) from the (visualization) tab, that's used to show the result of the experiment after loading the data from the (ArffLoader).

In addition to the previous components, The Attribute selection Filter (type of filters that can be chosen from the filter tab and used to specify the feature selection method that will be used with different classifiers have been used for each method) have been used with its default properties for the all methods except (method8) where its evaluator has been changed to(WrapperSubsetEval) and its (search) to (GeneticSearch) and (method 9)where its evaluator has been changed to (ClassifierSubsetEval) and its (search) to (GeneticSearch).

Method 1

As shown in figure (3), (BayesNet) classifier has been used, it represents a probabilistic relationship among a set of random variables graphically.

Method 2

As shown in figure (4), (LADTree) classifier has been used. It is a class that uses the "LogitBoost" strategy for generating a multi-class alternating decision tree.

Method 3

As shown in figure (5), (j48) classifier has been used. It represents a class for generating an un-pruned or a pruned C4.5 decision tree.

Method 4

As shown in figure (6), (NaiveBayesMultinominalUpdaeable) classifier has been used. It is a class for building and using a multinomial Naive Bayes classifier.

Method 5

As shown in figure (7), (FilteredClassifier) classifier has been used, which removes the necessity of filtering the data before the classifier can be trained.

Method 6

As shown in figure (8), (REPTree) classifier has been used, It represents a fast decision tree learner that builds a decision (regression) tree using information gain/variance and prunes it using reduced-error pruning.

Method 7

As shown in figure (9), (NaiveBayes) classifier has been used. (NaiveBayes) classifier algorithm based on Bayesian theorem and it is suited when the dimensionality of the inputs is high. In spite of its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

Method 8

As shown in figure (10), (ADTree) classifier has been used, which represents a class for generating an alternating decision tree.

Method 9

As shown in figure (11), the (ADTree) classifier have used with changing the properties of Attribute selection.

4. RESULTS

In this experiment, (9) different methods have been tested. These methods have been applied on each fold of the data set separately. Consequently, (9) results have been obtained. These results represent the accuracy of the Correctly Classified Instances that was obtained from each tried method. In this part, the average of the ten folds outputs has been calculated for each method separately. On the other hand, the time consuming by each fold of the data set was calculated while running the different (9) methods. The average of the time, including the time of the loading, consumed by the Ten folds set for each method have been calculated separately from other methods. The results have been shown in Table (3): Between these (9) methods, there are (3) methods that their classifiers are from (Bayes) family and (4) methods that their classifiers are from (Tree) family. Table (4) displays the (Bayes) family methods While Table (5) shows the (Tree) family methods, although they are already parts from the methods in table (3).

5. DISCUSSION

In this section, the alternative Classification techniques that have been used in the experiment will be compared in terms of the metrics reported in the previous section.

The main focus was given to the percentage of the correctly classified instances as well as the consuming time to run this experiment, which was greatly changed from the minimum (0:1:16) to maximum (0:17:09). The collected results were shown in table (4).

One of the good obtained results was from (FilteredClassifier) Algorithm. Although the obtained average of accuracy was not the best, but it could be considered as the best method has been run during this experiment. This is because it consumed only (0:01:37) to run it with good accuracy comparing to the consumption times by other methods with the default properties of the attribute selection.

Regarding the methods that their classifiers were chosen from (Bayes) family (table (4)), the most challenging point was in (NaiveBayes) algorithm, where the average of its accuracy was the best between the methods from the same family (Bayes) but it takes an average time about (0:17:09). This is a great amount of time comparing to (BayesNet) algorithm which spent less than a half of this amount of time with an average of accuracy a bit less than (NaiveBayes). As a result, it can be argued that it is not good to use (NaiveBayes) Classifier with the default properties of the attribute selection in experiments that is sensitive to the time consumption.

Comparing method (2) with method (6), it can be easily proved that it is not applicable to test more than one classifier from the (tree) family (with the default properties of the attribute selection) in the same experiment. This is because the average of the accuracy and the time consumption in both of them were approximately the same, although different classifiers have been used, but these classifiers were both from the same family (tree) as it is shown in table (5).

On the other hand, the components of (Method 8) and (Method 9) exactly the same, the difference was only in changing the properties of the attribute selection (evaluation). From table (5), it is clear that although they have exactly the same average of the accuracy of the Correctly Classified Instances; the time consumed by (Method 8) was approximately three times more than the time consumed by (Method 9).

Another challenging point was, the classifier (j48) have been used with the attribute selection (with changing the evaluator). It consumed more than (1:30:00) to run only one fold (fold 00) of the data set with accuracy not more than (71.4554). That is not a good result comparing either to the time it consumed or to the accuracy that was obtained from (Method 3) where the same classifier (j48) have been used with the attribute selection (with its default properties of evaluation and search).

The same operation tested again on the classifier (j48) with applying a second change on the evaluator of the attribute selection. It consumed more than (8 hours) to run only one fold (fold 00) of the data set.

From previous points, it can be assumed that changing the properties of the attribute selection (search and the evaluator) always leads to significant reduction in the time consumption.

As a result, if this experiment have been designed again, it is better to keep in mind the previous realizations in order to apply the classification process with high accuracy in less amount of time.

6. Figures and Tables

6.1. Figures:

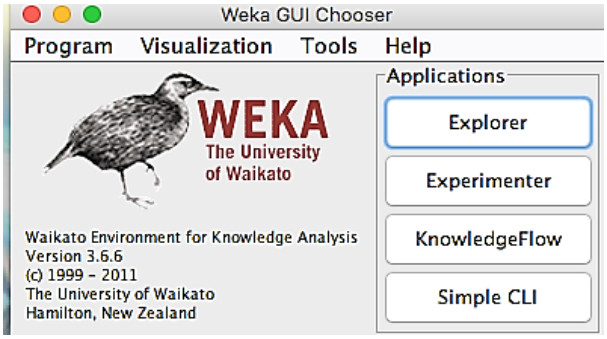

Figure 1: Testing data- load current (amperes)

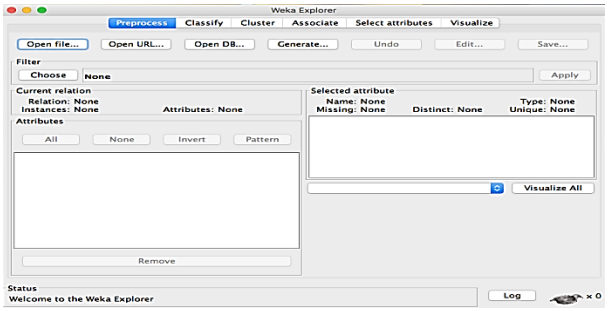

Figure 2: Weka Explorer

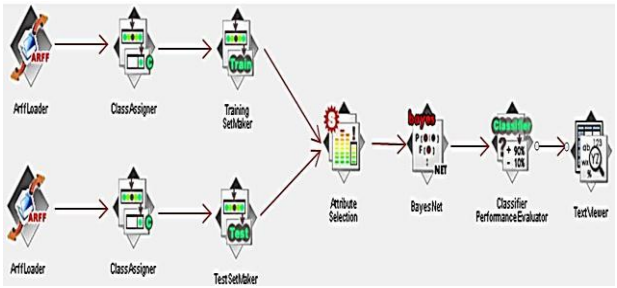

Figure 3: BayesNet Method

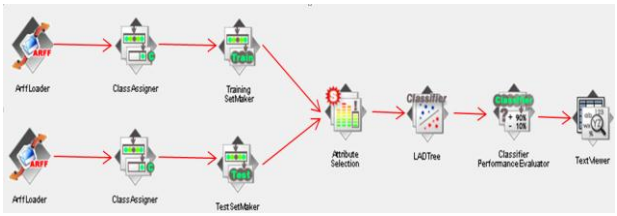

Figure 4: LADTree Method

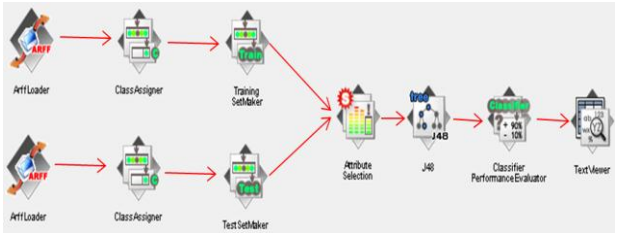

Figure 5: j48 Method

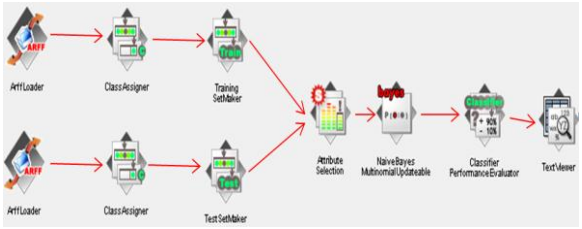

Figure 6: NaiveBayesMultinominalUpdaeable Method

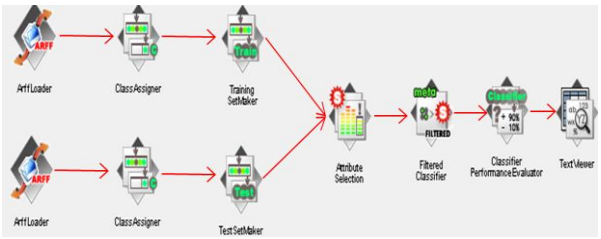

Figure 7: FilteredClassifier Method

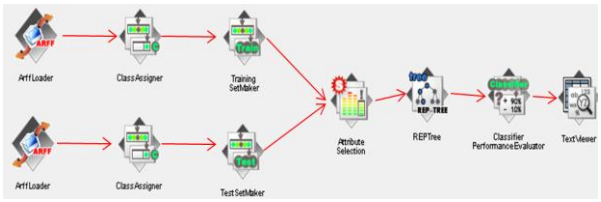

Figure 8: REPTree Method

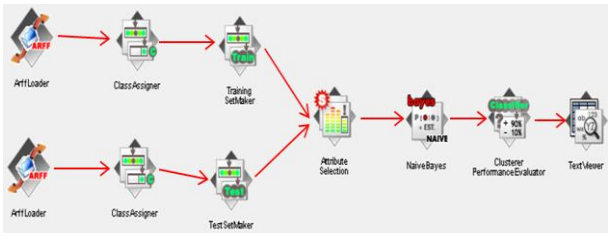

Figure 9 NaiveBayes Method

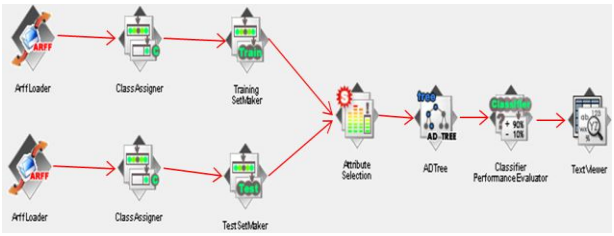

Figure 10 ADTree Method

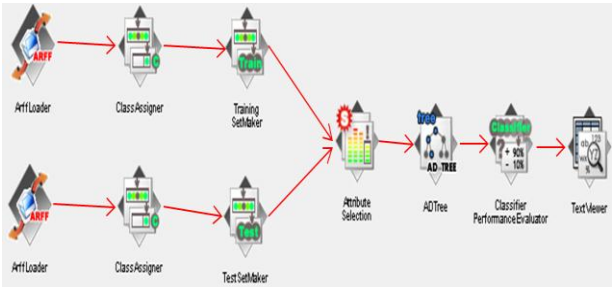

Figure 11: ADTree Method with Changing the Properties of Attribute Selection

6.2.Tables

Table 1: The Characteristic of the Training Set

CND	Number of instances	Number of attributes	type of attributes	class distibution	
				Name	amplitude
TrainFold00w4	234638	180	180 numeric	0	164080
				1	70558
TrainFold01w4	233012	180	180 numeric	0	163292
				1	69720
TrainFold02w4	232687	180	180 numeric	0	163001
				1	69686
TrainFold03w4	233573	180	180 numeric	0	160497
				1	73076
TrainFold04w4	232157	180	180 numeric	0	162799
				1	69358
TrainFold05w4	234255	180	180 numeric	0	164315
				1	69940
TrainFold06w4	234775	180	180 numeric	0	164122
				1	70653
TrainFold07w4	234294	180	180 numeric	0	161101
				1	73193
TrainFold08w4	233077	180	180 numeric	0	163290
				1	69787
TrainFold09w4	235654	180	180 numeric	0	164859
				1	70795

Table 2: The Characteristic of the Test Set

CND	Number of instances	Number of attributes	type of attributes	class distibution	
				Name	amplitude
TestFold00w4	22922	180	180 numeric	0	16379
				1	6543
TestFold01w4	24548	180	180 numeric	0	17169
				1	7379
TestFold02w4	24873	180	180 numeric	0	17458
				1	7415
TestFold03w4	23987	180	180 numeric	0	16617
				1	7370
TestFold04w4	25403	180	180 numeric	0	17660
				1	7743
TestFold05w4	23305	180	180 numeric	0	16144
				1	7161
TestFold06w4	22785	180	180 numeric	0	16337
				1	6448
TestFold07w4	23266	180	180 numeric	0	16013
				1	7253
TestFold08w4	24483	180	180 numeric	0	17169
				1	7314
TestFold09w4	21906	180	180 numeric	0	15600
				1	6306

Table (3): The Average of the Accuracy and the Time Consuming of the Correctly Classified Instances

Methods	Average	Time consuming
Mehtod1 (BayesNet)	76.25191	0:08:17
Mehtod2 (LADTree)	77.2657556	0:09:09
Mehtod3 (j48)	75.9483333	0:08:23
Mehtod4 (Naïve BayesMultinominalUpdaeable)	70.0338	0:10:20
Mehtod5 (FilteredClassifier)	76.53422	0:01:37
Mehtod6 (REPTree)	77.6177222	0:08:59
Mehtod7 (NaiveBayes)	77.0099222	0:17:09
Mehtod8 (ADTree)	70.1494556	0:03:42
Mehtod9 (ADTree Changing the Properties of Attribute Selection)	70.1494556	0:01:16

Table (4): Bayes Family Classifier Methods

Methods	Average	Time consuming
Mehtod1 (BayesNet)	76.25191	0:08:17
Mehtod4 (NaiveBayesMultinominal Updaeable)	70.0338	0:10:20
Mehtod7 (NaiveBayes)	77.0099222	0:17:09

Table (5): Tree Family Classifier Methods

Methods	Average	Time consuming
Mehtod2 (LADTree)	77.2657556	0:09:09
Mehtod3 (j48)	75.9483333	0:08:23
Mehtod6 (REPTree)	77.6177222	0:08:59
Mehtod8 (ADTree)	70.1494556	0:03:42
Mehtod9(ADTree/Changing the Properties of Attribute Selection)	70.1494556	0:01:16

7. CONCLUSION

The aim of this paper was to test the results of (9) different Classification methods applied on one a large scale dataset. The main focus was given in analysing and learning the Classification techniques rather than the data set itself. These methods were differs from each other in terms of the type of the classifier and/or the properties of the attribute selections that have used. From the obtained results, it became obvious that in this case of a large-scale bioinformatics dataset that comes from the protein structure prediction field, we can conclude the following points:

- It is not applicable to test more than one classifier from the (tree) family, with the default properties of the attribute selection, in the same experiment, as the result will be approximately the same
- Comparing to the other classifiers in the (Bayes) family, it is not preferred to use (NaiveBayes) Classifier with the default properties of the attribute selection in experiments care about the time consuming.
- Varying the attributes of the feature selection (evaluator and search) always affect the result.
- Among the seventh classifiers algorithms that have been used with the default properties of the attribute selection, the (FilteredClassifier) algorithm has better classification accuracy comparing to consumption time over and above compared algorithms.

Finally, Fine-tuning the experiment platform changing the parameters of the attribute selections should be prioritized for more accurate results in any experiments further.

8. REFERENCE

[1] AL-Nabi, Luqman Delveen, and Shukri Shereen Ahmed. "Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)." Computer Engineering and Intelligent Systems 4, no. 8: pp.18-27 (2013).

[2] Angus-Hill, et al. "A Rsc3/Rsc30 zinc cluster dimer reveals novel roles for the chromatin remodeler RSC in gene expression and cell cycle control." Molecular cell 7, no. 4: pp.741-751(2001).

[3] Bergmann, Sven, Ihmels Jan, and Barkai Naama. "Iterative signature algorithm for the analysis of large-scale gene expression data." Physical review E 67, no. 3: pp.031902 (2003).

[4] Bhavsar, H., and A. Ganatra. "A comparative study of training algorithms for supervised machine learning." International Journal of Soft Computing and Engineering (IJSCE) 2, no. 4: pp.2231-2307 (2012).

[5] Chauhan, R., H. Kaur, and M. A. Alam. "Data clustering method for discovering clusters in spatial cancer databases." International Journal of Computer Applications 0975–8887 (2010).

[6] David, S. K., A. T. Saeb, and K. Al Rubeaan. "Comparative Analysis of Data Mining tools and classification Techniques using WEKA in Medical Bioinformatics." Computer Engineering and Intelligent Systems 4, no. 13: pp.28-38 (2013).

[7] Dueck, D., D. Q. Morris, and J. B. Frey. "Multi-way clustering of microarray data using probabilistic sparse matrix factorizatio."

Bioinformatics 21, no. suppl 1: pp.i144-i151 (2005).
[8] Eisen, M. B., P. T. Spellman, P. O. Brown, and Botst. "Cluster analysis and display of genome-wide expression patterns." Proceedings of the National Academy of Sciences 95, no. 25: pp.14863-14868 (1998).
[9] Erica, C., and H. Falk. "Using Blackbox Algorithms Such as TreeNet and Random Forests for Data-Mining and for Finding Meaningful." Information science reference,: pp. 65-84 (2009).
[10] Everitt, S. B., Landau Sabine, and Leese Morven. Cluster Analysis. fourth. London: Arnold, (2004).
[11] Fayyad, U., and S. Paul. "Data mining and KDD: Promise and challenges." Future generation computer systems 13, no. 2-3: pp.99-115 (1997).
[12] Frank, E., M. Hall , L. Trigg, G. Holmes, and I. H. Witten. "Data mining in bioinformatics using Weka." Bioinformatics 20, no. 15: pp.2479-2481 (2004).
[13] Freitas, A. A. "Data mining and knowledge discovery with evolutionary algorithms." Springer Science & Business Media, (2013).
[14] Guerra, L., M. McGarry, V. Robles, C. Bielza, P. Larrañaga, and R. Yuste. "Comparison between supervised. And unsupervised classifications of neuronal cell types: A case study." Developmental neurobiology 71, no. 1: pp. 71-82 (2011).
[15] Huttenhower, C., M. Hibbs, C. Myers, and Troyansk. "A scalable method for integration and functional analysis ofmultiple microarray datasets." Bioinformatics 22, no. 23: pp.2890-2897 (2006).
[16] John, G. H., and P. Langley. "Estimating continuous distributions in Bayesian classifiers." In Proceedings of the Eleventh conference on Uncertainty in artificial intelligence (Morgan Kaufmann Publishers Inc),: pp. 338-345 (1995).
[17] Kifaya, S. Qaddoum. "Mining Student Evolution Using Associative Classification and Clustering." Communications of the IBIMA 11, no. 1943-7765: pp. 19-25 (2009).
[18] Kretschmann, E., W. Fleischmann, and R. Apweiler. "Automatic rule generation for protein annotation with the C4. 5 data mining algorithm applied on SWISS-PROT." Bioinformatics 17, no. 10: pp.920-926 (2001).
[19] Li, J., and L. Wong. "Identifying good diagnostic gene groups from gene expression profiles using the. Concept of emerging patterns." Bioinformatics 18, no. 5: pp. 725-734 (2002).
[20] Luscombe, N. M., D. Greenbaum, and M. Gerstein. "What is bioinformatics? An introduction and overview." Yearbook of Medical Informatics 1, no. (83-100): p.2 (2001).
[21] Pavlidis, P., J. Weston, J. Cai, and W. S. Noble. "Learning gene functional classifications from multiple data types." Journal of computational biology 9, no. 2: pp.401-411 (2002).
[22] Pi, Jiaxiong, Yong Shi, and Z. Chen. "From similarity retrieval to cluster analysis: The case of R*-trees." Computational Intelligence and Data Mining,: pp. 524-529 (2007).
[23] Rahman, R. M., and F. Afroz. "Comparison of various classification techniques using different data mining tools for diabetes diagnosis." Journal of Software Engineering and Applications 6, no. 03: p.85 (2013).
[24] Schreiber, A. W., and U. Baumann. "A framework for gene expression analysis." Bioinformatics 23, no. 2: pp.191-197 (2007).
[25] Tan, A. X., V. L. Liu, M. Kantarcioglu, and Thurais. "A comparison of approaches for large-scale data mining." Technical Report UTDCS-24–10, Tech. Rep., (2010).
[26] Thakur, R., and A.R. Mahajan. "Preprocessing and Classification of Data Analysis in Institutional System using Weka." International Journal of Computer Applications 112, no. 6 (2015).
[27] Tobler, J.B., M.N. Molla, E.F. Nuwaysir, R.D. Green, and J.W. Shavlik. "Evaluating machine learning approaches for aiding probe selection for gene-expression arrays." Bioinformatics 18, no. (suppl 1): pp.S164-S171 (2002).
[28] Troyanskaya, O.G., K. Dolinski, A.B. Owen, R.B. Altman, and D. Botstein. "A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)." Proceedings of the National Academy of Sciences 100, no. 14: pp.8348-8353 (2003).
[29] Yang, H. C., A. Dasdan, R. L Hsiao, and D. S. Parker. "Map-reduce-merge: simplified relational data processing on large clusters." In Proceedings of the 2007 ACM SIGMOD international conference on Management of data,: pp. 1029-1040 (2007).
[30] Yoo, I., et al. "Data mining in healthcare and biomedicine: a survey of the literature." Journal of medical systems 36, no. 4: pp.2431-2448 (2012).

9. Biography

1- **Nigar M. Shafiq Surameery**
https://scholar.google.com/citations?hl=en&user=3Kp8XWgAAAAJ

2- **Dana Lateef Hussein**

https://scholar.google.co.uk/citations?user=NZv2E-sAAAAJ&hl=en