An Integrated Framework for Knowledge Extraction using Clustering and Classification

Thesis Summary

submitted to

The Maharaja Sayajirao University of Baroda

For the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

by

Mrs. Subodhini Gupta

Research Guide

Dr. Anjali Ganesh Jivani

Department of Computer Science and Engineering
Faculty of Technology and Engineering

The Maharaja Sayajirao University of Baroda

Vadodara 390 001

June 2020

Table of Contents

Abstract ..... **vi**
Table of Contents ..... **viii**
List of Figures ..... **xi**
List of Tables ..... **xiii**
Chapter-1: Introduction **1**
1.1 Overview of Data Mining 1
1.2 The Data Mining Process 4
1.2.1 Data Selection 5
1.2.2 Data Preprocessing 6
1.2.3 Data Mining 7
1.2.4 Pattern Evaluation 7
1.2.5 Knowledge Presentation 7
1.2.6 Interpretation 7
1.2.7 Use of discovered knowledge 7
1.3 Data mining task 8
1.3.1 Predictive mining task 8
1.3.2 Descriptive task 9
1.4 Models for Data mining Tasks 10
1.5 Data Mining Tools 11
1.5.1 WEKA 11
1.5.2 ORANGE 12
1.6 Problem Statement and Objectives 13
1.6.1 Problem statement 13
1.6.2 Research Objectives 13
1.7 Research Contribution 13
1.8 Organization of the thesis 16
Chapter- 2: Literature Review **18**
2.1 Data Mining Techniques 18
2.2 Data Mining Tools 22
2.3 Improve the Performance of Classification Algorithm 30
2.4 Imbalanced Data 33
2.5 Feature Selection 35
2.6 Research gap 37
2.7 Summary 38

Chapter-3: Clustering & Classification 39
3.1 Classification 39
3.2 Classifiers Performance measuring Parameters 40
3.3 Classification techniques 42
3.3.1 Regression 44
3.3.2 Bayesian Classification 44
3.3.3 Decision Tree 45
3.3.4 The KNN (K-Nearest Neighbor) 48
3.3.5 Support Vector Machines 48
3.3.6 NN supervised learning 49
3.3.7 Random Forests 51
3.4 Clustering 53
3.4.1 Clustering Techniques 54
3.5 K-means algorithm 55
3.6 Distance metrics 57
3.7 Summary 58
Chapter-4: A Cluster-based solution for Imbalance Data 59
4.1 Introduction 59
4.2 Preliminaries and basic definitions 60
4.2.1 Imbalanced Data 60
4.2.2 Between - class imbalance & within- class Imbalance 62
4.2.3 Imbalance Ratio 63
4.2.4 Degree of imbalance distribution 63
4.2.5 False positive and false negative 63
4.2.6 Clustering 64
4.2.7 SVM Classifier 64
4.3 Methods of handling Imbalanced Data 65
4.4 Experimental investigations 67
4.4.1 Datasets 67
4.4.2 Experiment Setting 67
4.5 Proposed Cluster Based Under-sampling 69
4.6 Methodology 70
4.7 Summary 84
Chapter-5: Feature Selection through Clustering to Classify High Dimensional Data 85
5.1 Introduction 85
5.2 Introduction about Data set 88

5.3 Preliminaries and basic definitions 89
5.3.1 SVM Classifier 89
5.3.2 RELIEF Feature Selection Approach 89
5.3.3 Info-Gain Feature Selection Approach 90
5.5 Methodology 90
5.6 Performance analysis of the Generated Model 96
5.7 Summary 104
Chapter-6: A Hybrid Model to Enhance the Performance of a Classifier 106
6.1 Introduction 106
6.2 Classification 107
6.2.1 SVM 107
6.3 Clustering 108
6.3.1 K-Means 108
6.4 Models for Data mining Tasks 109
6.4.1 WEKA 111
6.5 Datasets Description 112
6.6 Methodology 113
6.6.1 Data Preparation 113
6.6.2 Clusters Building 113
6.2.3 Building the classification Models: 113
6.3 Algorithm 115
6.4 Performance analysis of the generated model 117
6.5 Summary 123
Chapter-7: Conclusion and Future Work 124
7.1 Conclusion 124
7.2 Future Enhancements 125
Publications 127
References 128

Table of Contents of Thesis Summary

Sr. No.	Topics	Page No.
1.1	Introduction	1
1.2	Research Gap	1
1.3	Research Objectives	2
1.4	Research Methodology	2
1.5	Layout of the Thesis	5
1.6	Conclusion	6
1.7	Recommendations	6
1.8	Bibliography	7

Thesis Summary

1.1 Introduction

The overall objective of this study is to develop an accurate, simple, and understandable model by combining supervised (Classification) and unsupervised (Clustering) learning methods. This study has proved that the integrated model does not only improves the accuracy of the classifier but it also handles some challenging issues such as Imbalanced, Noisy, incomplete, and high dimensional datasets, that have existed in Data mining areas for a long time.

1.2 Research Gap

There are several machine learning algorithms available to perform a specific task. A specific algorithm works on a specific type of data and fails to retain its performance with a different set of data. Sometimes it becomes difficult to choose this combination. Furthermore, there are many issues related to data itself which restricts good algorithms to perform well. It is also observed in the literature survey that there are several algorithms proposed by domain experts to deal with the aforementioned issues associated with data. For example, if the data is imbalanced one needs to learn a different set of algorithms that deals with imbalanced data same for high dimension, noisy, or data having more missing values one has to learn several algorithms to identify the best one. This problem is addressed by Manuel Fernandez-Delgado et. al. (2014) in his paper —Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" They quoted that —A researcher may not be able to use

classifiers arising from areas in which he/she is not an expert, being often limited to use the methods within his/her domain of expertise".

There is not any single algorithm to solve these problems implicit in real-life data. To fill this gap, I have proposed an integrated framework of clustering and classification as a generic solution to deal with these challenges present in the data. In this approach, one has to learn only K means algorithm for clustering and SVM classifier for the classification task to deal with every kind of data

1.3 Research Objectives

The detailed research objectives are to integrate the methods that can serve the following purposes.

I. Balancing Imbalance data for traditional Classifiers

II. Improve the performance of the traditional classifiers

III. Feature compression & extraction technique

Three models are presented and implemented that offer a generic and logical solution to meet up the research objectives.

1.4 Research Methodology

1.4.1 Model-1

The first proposed model provides a Cluster-based approach using an under-sampling solution to balance the imbalanced data. A study for binary class distribution on 12 data sets openly available in UCI machine learning repository with different degrees of imbalance nature has been conducted. The proposed framework is capable in giving a three-fold solution. Firstly, it balances imbalanced data using the K-means

clustering approach. The main purpose of this approach is to selectively discard the majority instances from the dataset to make the distribution balanced and can be applied to any traditional classifier. Secondly, it can handle between-class imbalance distribution and within-class distribution. Thirdly, it can handle the different degrees of imbalance distribution.

Research findings: The performance of the algorithm is considerably good with highly imbalanced datasets. It has reported an improved value of f-measure for approx all the dataset except the wine quality white dataset because the data sets contain many missing values.

1.4.3 Model-2

The second proposed model is applicable as a Feature compression technique to build a better feature space because it can solve several machine learning problems. High dimensional data generally diminish the accuracy and efficiency of Data Mining algorithms. The higher the dimensionality, the higher the computation cost involved in processing. Irrelevant features may exert undue burden on classification evaluation parameters and increases the time and resources needed to build the model. The experiments were conducted on nine benchmark datasets taken from UCI machine repository with diverse degrees of dimensionality. The comparative analysis of the proposed approach with a standard correlation-based Feature Selection approach (RELIEF and Info-Gain Approach) was presented on the datasets. The results revealed that it reduces dimensionality, complexity, and computation time, and increases comprehensibility and the overall performance of classification algorithms. Therefore,

the proposed model first identifies the relevant features that may lead to accurate results.

Research findings: The performance of the proposed algorithm is better compare to relief and info gain state-of-the-art feature selection methods with every percentage of features sets, at 60% of features the performance of the proposed model is very impressive.

1.4.2 Model-3

The Third proposed model is to improve the performance of any classifier using a hybrid model (prior clustering to classification). The research experiments observed that the hybrid model is more refined and accurate than a single individual classifier. The objective is to utilize the strength of one method to complement the weaknesses of another. If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as a hybrid model. The experiment has been conducted on 13 benchmark datasets taken from UCI machine learning repository. The results revealed that this integration process generates a more precise and accurate model. The major classifier's accuracy gets affected positively when supervised and unsupervised learning methods are combined.

Research findings: It is identified that the accuracy of the proposed model is reasonably high with all set of values for K. And K=3 is the most appropriate value in all cases except for haberman dataset where K=5 is the best value because it is less dimensional.

The developed models have been published in Conference Proceedings /International Journals.

1.5 Layout out of the Thesis

The contributions from this study have been presented in the chapters as follows:

Chapter-1: This chapter contains the general introduction of the broad area of the research it contains an overview of Data mining, Data mining application areas, Data mining process, Data mining tasks, models for Data mining tasks, classification and clustering techniques and tools.

Chapter-2: An extensive literature review has been accomplished and presented in this chapter.

Chapter -3: A detailed description of Clustering and Classification and the techniques used to perform these tasks are presented in this chapter.

Chapter-4: This chapter demonstrates the application of this proposed integrated approach in handling Imbalance class distribution in real-world applications which is one of the top challenging problems in data mining.

Chapter -5: This Chapter contains the third proposed model, which is applicable as a Feature compression technique to build a better feature space because it can solve several machine-learning problems.

Chapter-6: This chapter proposed an integrated approach, which is applicable in various aspects of data mining in this chapter model, is deployed to use clustering as a preprocessing process for classification. It reduces the training data set size and its

dimensionality by dividing the whole Dataset into smaller sub-sets leads to a smaller and less complicated classification task becomes quicker and easier to solve.

Chapter- 7 Conclusion & Future work

Chapter- 8 References.

1.6 Conclusion

The research work carried out developed three models.

The first model is capable of handling between-class Imbalance distribution and within-class distribution .it can also handle different degrees of imbalanced distribution and finally, it balances the imbalanced data and that can be applied to any traditional classifier. The proposed model is simple yet effective in order to classify the Imbalanced distribution of Data.

The second model is applicable as a Feature selection technique to build a better feature space. It is observed that with 60% of features the performance of the proposed model is very impressive.

In the Third model, it is observed that the hybrid model is more refined and accurate than a single individual classifier. If we are interested in the best possible classification accuracy, it might be difficult to find a single classifier that performs as good as a hybrid model. The results evidently revealed that this integration process generates a more precise and accurate model.

1.7 Recommendation: The Framework can be used to deal with more real-world data problems such as unlabelled data, unstructured data, and Textual data.

1.8 Bibliography

Abdulla, H.W. et al. 2011. A Comparison Study between Data Mining Tools over some Classification Methods. In *Proceedings of International Conference of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, pp. 18-26.

Ahmed, M.S., and Khan, L. 2009. SISC: A Text Classification Approach Using Semi-Supervised Subspace Clustering 2009, IEEE International Conference on Data Mining Workshops.

Ali, A., Shamsuddin, S.M. and Ralescu, A.L. 2015. Classification with class imbalance problem: A Review.International Journal of Advance Soft Compu. Appl, Vol. 7, No. 3, ISSN 2074-8523.

Anooj, P.K. 2012. Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. *Journal of computer sciences, vol.24*, pp. 27- 40.

Arora, R., Suman 2012.Comparative Analysis of Classification Algorithms on Different Datasets using WEKA.International Journal of Computer Applications, Volume 54.

Arunadevi, J., Ramya, S., Raja, R.M. 2018. A study of classification algorithms using Rapidminer. *International Journal of Pure and Applied Mathematics* Volume 119 No. 12.

Asha, T., Natarajan, S., Murthy, K. N.B. 2014.A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification.

Bhargavi, P., Jyothi, S. 2011. Soil Classification Using Data Mining Techniques: A Comparative Study International Journal of Engineering Trends and Technology.

Bielza, C. and Larranaga, P. 2014.Discrete Bayesian network classifiers: a survey, ACM Computing Surveys, vol. 47.

Boudour, M., & Hellal, A. 2005.The combined use of supervised and unsupervised learning for power system dynamic security mapping. Engineering Applications of Artificial Intelligence, 18, 673–683.

Breault, J.L.2001. Data Mining Diabetic Databases: Are rough Sets a Useful Addition?" http://www.galaxy.gmu.edu/interface/I01/I2001Proceedings/Jbreault.

Breiman, L.,Friedman, J. H.,Olshen, R. A., Stone, C.J. 1984. Classification and Regression Trees, Wadsworth Books.

Burges, C.J. 1998. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery 2, 121–167.

Butterworth, R., Gregory, P., Dan, A. 2005. On Feature Selection through Clustering. Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05).

Chandrashekar, G., Sahin, F. 2014. A survey on feature selection methods. Computers & Electrical Engineering, Volume 40, Issue 1, 2014, Pages 16-28, ISSN 0045-7906.

Chawla, N. 2005. Data Mining for Imbalanced Datasets: An Overview. 10.1007/0-387-25465-X_40.

Chawla, N. V., Japkowicz, N., and Kokz, A. 2004. Learning from Imbalanced Datasets. SIGKDD Special Issue.

Chawla, N.V., Bowyer, K.W., Hall, L.O.and Kegelmeyer, W.P.2002. SMOTE Synthetic minority over-sampling technique. Journal of artificial intelligence research, Vol.16 pp. 321-357, 2002.

Chawla, N.V., Japkowicz, N., and Kolcz, A. 2003. Learning from Imbalanced Data Sets. Proc. Int'l Conf. Machine Learning.

Chen, F., Deng, P., Wan, J., Zhang, D, Athanasios V., and Ron, X. 2015. Review Article Data Mining for the Internet of Things: Literature Review and Challenges. *Hindawi Publishing Corporation International Journal of Distributed Sensor Networks*, Article ID 431047.

Chitra, R. 2013. Review of heart disease prediction system using data mining and hybrid intelligent techniques. Journal on soft computing, Volume: 03, Issue: 04.

Chujai, P., Chomboon, K., Chaiyakhan, K., Kerdprasop, K., Kerdprasop, N.2017. A Cluster-Based Classification of Imbalanced Data with Overlapping Regions Between Classes.*Proceedings of the International MultiConference of Engineers and Computer Scientists Vol I, IMECS , Hong Kong*.

Çıgsar, B. and Deniz, U. 2019. Comparison of Data Mining Classification AlgorithmsDetermining the Default Risk. *Hindawi Scientific Programming.*

Cover, T., and Hart, P. 1967.Nearest neighbor pattern classification.IEEE transactions on information theory, vol 13, pp.21-27.

Covoes, T., F., Hruschka, E.R. 2011. Towards improving cluster-based feature selection with a simplified silhouette, filter. *Information Sciences 181 3766–3782.*

Davis, D.N., Rahman M. 2013. Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing, Vol. 3, No. 2.*

Davis, J., and Goadrich, M. 2006.The relationship between Precision-Recall and ROC Curves", 23 rd International Conference on Machine Learning, Pittsburgh, PA.

Delen, D., Walker, G., & Kadam, A. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, 34, 113-127.

Demšar, J., Zupan, B, Leban, G., Curk, T. 2004.Orange: From Experimental Machine Learning to Interactive Data Mining, Knowledge Discovery in Databases: PKDD".

Deng, X., Liu, Q., Deng, Y. Mahadevan, S. 2016. An improved method to construct basic probability assignment based on the confusion matrix for classification problems. Information Sciences, Volumes 340–341, Pages 250-261, ISSN 0020-0255.

Dietterich T.G. 2000. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. *Lecture Notes in Computer Science, vol 1857*, Springer, Berlin, Heidelberg.

Dimitoglou, G., Adams, J. A. and Jim, C. M. 2012. Comparison of the C4.5 and a Nave Bayes Classifier for the Prediction of Lung Cancer Survivability. Journal of Computing, Vol. 4, No. 2, pp. 1-9.

Drown, D. J., Khoshgoftaar, T. M., and Seliya, N. 2009. Evolutionary Sampling and Software Quality Modeling of High-Assurance Systems. IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, Vol. 39, No. 5.

Duda, R. O., Hart, P.E., and Stork, D.G. 2001. Pattern Classification. John Wiley & Sons Inc., USA.

Dunham, M. H. 2006. Data mining introductory and advanced topics. Pearson Education India .

Durairaj, M., Ranjani, V. 2013 .Data Mining Applications, In Healthcare Sector ,International Journal of Scientific & Technology Research, ISSN 2277-8616 29.

Elder, J. and Goldberg, R. 2001. Image Editing in the Contour Domain. *IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 21, no. 03*, pp. 291-296.

Elkan, C. 2004. Clustering with k-means: Faster, smarter, cheaper .Keynote talk at Workshop on Clustering High-Dimensional Data, SIAM International Conference on Data Mining.

Ester, M., Kriegel, H, Sander, J. Xu, X. 1996. Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise From",KDD AAAI (www.aaai.org).

Fahad, A. 2014. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Transactions on Emerging Topics in Computing, 267-279.

Fahy, C., Yang, S. 2019. Dynamic Feature Selection for Clustering High-Dimensional Data Streams .Special section on AI driven big data processing: Theory, methodology and applications IEEE access Volume.

Fayyad U., Piatetsky-Shapiro, G. and Smyth, P. 1996 (a) .The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39.11, 27-34.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. 1996 (b) .From Data Mining to Knowledge Discovery: An Overview. Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, pp.1-34.

Fayyad, U. and Smyth, P. 1996. Advances in knowledge Discovery and Data Mining. AAAI/MIT press, Menlo park CA.

Feng C, F., Deng, P., Wan, J., Zhang, D., Vasilakos, A.V., and Rong, X. 2015. Data Mining for the Internet of Things: Literature Review and Challenges. *Hindawi Publishing Corporation International Journal of Distributed Sensor Networks.*

Fernández, A., Río, S. D., Chawla, N. V., Herrera, F.,2016. An insight into imbalanced Big Data classification: outcomes and challenges .Complex Intell. Syst Springer.

Fodor, I. K . 2002. A survey of dimension reduction techniques. Department of Energy by the University of California. 1-7.

Frank, E.., Mark, H., Len, T., Holmes, G.,Ian, W. 2004. Data mining in bioinformatics using Weka "Bioinformatics Applications, pages 2479–2481.

Frawley, W.J., Shapiro, G.P., and Matheus, C. J., Fayyad, U., Smyth, P. 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework in KDD-96 Proceedings.

Fu, Y. 1997. Data Mining, IEEE potentials, vol.16,pp.18-20.

Guha, S., Rastogi, R. and Shim, K. 1998 .Cure: An efficient clustering algorithm for large databases. Proc. of the ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98), pp. 73-84.

Gupta S., Parekh B. S., Jivani A. (2019) A Hybrid Model of Clustering and Classification to Enhance the Performance of a Classifier. In: Luhach A., Jat D., Hawari K., Gao XZ., Lingras P. (eds) Advanced Informatics for Computing Research. ICAICR 2019. Communications in Computer and Information Science, vol 1076. Springer, Singapore.

Haixianga, G., Yijinga, L., Shang, J., Mingyuna, G.,Yuanyuea, H., Bing, G. 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73 (2017) 220–239.

Hall, M., Frank, E., Holmes, G. ,Pfahringer, B. ,Reutemann, P., Ian,W. , The WEKA Data Mining Software: An Update", SIGKDD Explorations Volume 11, Issue 1.

Han, J. and Kamber, M. 2001. Data Mining: Concepts and Techniques. San Francisco, Morgan Kauffmann Publishers.

Han, J., Rodriguze, J.C., Beheshti, M. 2008.Diabetes data analysis and prediction model discovery-using rapidminer. *Second International Conference on Future Generation Communication and Networking, pp. 96-99*.

Hardin, J. M., & Chhieng, D. C. 2007. Data Mining and Clinical Decision Support Systems . In Hannah, Health Informatics, formerly.

Harding J. A, Shahbaz M. M, Srinivas, J., Kusiak A. A. 2005. Data Mining in Manufacturing: A Review. ASME. Manuf. Sci. Eng. 128(4):969-976. doi:10.1115/1.2194554.

Haykin, S., 2009 .Neural Networks and Learning Machines .Third Edition by Pearson Education.

He, H. and Garcia, E. A. 2009. Learning from Imbalanced Data IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9.

Hearst M.A. 1998. Support Vector Machines .IEEE intelligent systems.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt J. and Scholkopf, B. 1998. Support vector machines . *IEEE Intelligent Systems and their Applications, vol. 13, pp. 18-28.*

Höppner, F., Klawonn, F., Kruse, R., and Runkler, T. 2000 Fuzzy Cluster Analysis .Chichester John Wiley & Sons .

Huang, A. 2008.Similarity Measures for Text Document Clustering .The New Zealand Computer Science Research Student Conference.

Huang, J., Lu, J. and Ling, C. X. 2003 .Comparing Nave Bayes, Decision Trees, and SVM with AUC and Accuracy .Proceedings of Third IEEE International Conference on Data Mining, 19-22, pp. 553-556. doi:10.1109/ICDM.2003.1250975.

Hwang, W. J. and Wen, K. W. 1998 .Fast KNN classification algorithm based on partial distance search, Electronics Letters, vol. 34, no. 21, pp. 2062–2063.

Ismia, D. P., Shireen P, Murintoc. 2016. K-means clustering based filter feature selection on high dimensional data. International Journal of Advances in Intelligent Informatics ISSN: 2442-6571, pp. 38-45.

J. R. Quinlan, "Induction of decision trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986.

Jain A. K. 2010. Data clustering 50 years beyond K-means. *Pattern Recognition Letters* pp. 651–666.

Jain A. K., Murthy M. N, and Flynn P. J. 1999 . Data clustering a review .ACM Comput. Surv. 31, 264-323.

Jain, A.K., and Dubes, R. C. 1988. Algorithms for Clustering Data.,Prentice Hall, New Jersey.

Jakkula, V. 2006 .Tutorial on Support Vector Machine (SVM) .*School of EECS, Washington State University*, *37*.

Jannet, P. W. et al. 2012 .Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News .International Journal of Computer Applications Volume 50 – No.11, July 2012.

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-A1'2000): Special Track on Inductive Learning, Las Vegas, Nevada.

Jason, D. M., Lawrence, S.R., Teevanteevan, J., David R. 2003. Tackling the Poor Assumptions of Naive Bayes Text Classifiers. *Twentieth International Conference on Machine Learning (ICML-2003), Washington DC.*

Jeng-Shyang, P., Yu-Long, Q. and Sheng-He S.. 2004. Fast k-nearest neighbors classification algorithm, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. 87, no. 4, pp. 961–963.

Johnson, J.M., Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. *J Big Data 6, 27 . https://doi.org/10.1186/s40537-019-0192-5*.

Julio F. N., Frenk, C.S. Simon D. M. 1997. White a Universal Density Profile from Hierarchical Clustering", The Astrophysical Journal.

Karegowda, A. G. et al. 2012 .Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012.

Karegowda, J., Punya, V., Manjunath, A.S. 2012. Rule based classification for diabetic patients using cascaded k-means and decision tree c4.5. *International Journal of computer applications, Pp. 45–12.*

Kaufman, L. and Rousseeuw,P. Z. 1990 .Finding groups in data: An introduction to cluster analysis, New York: John Wiley & Sons.

Kaufmann, M. 1993. C4. 5 : *Programs for Machine Learning, vol. 1.*

Kaur, G. Chhabra, A. 2014. Improved J48 Classification Algorithm for the Prediction of Diabetes International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22.

Kesavaraj, G. & Surya, S. 2013. A study on classification techniques in data mining", 4th International Conference on Computing, Communications and Networking Technologies, p.p 1-7.

King, R. D., Feng, C., & Sutherland, A. 1995. Comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence, 9(3), 289- 333.

Kira, K., Rendell, L. A. 1992 . The feature selection problem: traditional methods and a new algorithm .AAAI, vol. 2, pp. 129–134.

Kotsiantis, B. S. 2007. Supervised Machine learning: A Review of classification Techniques. *Emerging Artificial intelligence Applications in Computer Engineering I. Maglogiannis* , IOS Press .

Kotsiantis, S., Kanellopoulos, D., Pintelas, P. 2006. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, Vol.30.

Kumar, A. and Kumar A. 2011. Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques.IEEE, pp: 161-165.

Kyriakopoulou, A. 2008. Text Classification Aided by Clustering: A Literature Review, Tools in Artificial Intelligence, Paula Fritzsche (Ed.), ISBN: 978-953-7619-03-9.

Kyriakopoulou, A. and Kalamboukis, T. 2008 .Combining clustering with Classification for Spam Detection in Social Bookmarking Systems. *ECML/PKDD Discovery Challenge.*

Ladla, L. and Deepa, T. 2011 . Feature Selection Methods And Algorithms . International Journal on Computer Science and Engineering (IJCSE), vol.3 (5), pp. 1787-1797, 2011.

Laopracha, N., Sunat, K., Chiewchanwattana, S. 2019 .A Novel Feature Selection in Vehicle Detection Through the Selection of Dominant Patterns of Histograms of Oriented Gradients (DPHOG). Access IEEE, vol. 7, pp. 20894-20919.

Larose, D. T. 2005. K-nearest neighbor algorithm in Discovering Knowledge in Data: An Introduction to Data Mining", pp. 90–106, John Wiley & Sons.

Lee C, Lee G. 2006. Information gain and divergence based feature selection for machine learning-based text categorization .Information Processing & Management.155-65.

Lee, C., Lee, G. G. 2006. Information gain and divergence-based feature selection for machine learning-based text categorization . Information processing & management volume 42, issue 1, January 2006, Pages 155-165.

Lee, J., and Siau, K 2001. A review of data mining techniques , *Industrial Management & Data Systems,* Vol. 101, pp. 41-46.

Lei, Y. and Liu, H. 2003. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC.

Li, D. C., Liu, C. W., Hu, S. C. 2011. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets , Artificial Intelligence in Medicine, 52, 45–52.

Liao, W. Liu, Y., Choudhary, A. 2004. A Grid-based Clustering Algorithm using Adaptive Mesh Refinement appears in the 7th Workshop on Mining Scientific and Engineering Datasets.

Lim, T., Loh, W., and Shih, Y. 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms , Machine Learning, 40, 203–228.

Little, M., and McSharry, P. 2008. Suitability of dysphonia measurements for tele-monitoring of Parkinson's disease. *Nature Precedings. 1-27.*

Liu, W., Liu, S., Gu, Q., Chen, X., Che, D. 2015.FECS: A Cluster based Feature Selection Method for Software Fault Prediction with Noises. IEEE 39th Annual International Computers, Software & Applications Conference.

Liu, X., Wu, J. and Zhou, Z. Under sampling for Class-Imbalance Learning. IEEE transactions on systems, man and cybernetics – part b.

Liu, Y. et al. 2011. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets, *Information Processing & Management, vol. 47, no. 4, pp. 617-631.*

Loizou, G. and Maybank, S.J. 1987 .The nearest neighbor and the bayes error rates .IEEE transactions on pattern analysis and machine learning, PP.254-263.

López, M.I et al Classification via clustering for predicting final marks based on student participation in forums, Proceedings of the 5th International Conference on Educational Data Mining.

Lopez, V., Fernandez, A., Garcia , S.,Palade , V., Herrera ,F.2013. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics Information Sciences 250 (2013) 113–141.

Luukka, P. 2011 .Feature selection using fuzzy entropy measures with similarity classifier", Expert Systems with Applications, 38, 4600–4607.

Manikandan, V., and Latha, S. 2013. Predicting the analysis of heart disease symptoms using medicinal data mining methods .*International Journal of Advanced Computer theory and Engineering, vol. 2, pp.46-51*.

Manzalawy, Y. and Honavar, V. 2005 .LSVM: Integrating LibSVM into Weka Environment, [http://www.cs.iastate.edu/~yasser/wlsvm].

Mariscal, G., Marbán, Ó. , Fernández, C. 2010. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, pp. 137-166.

Masand, B. and Shapiro, G. P. 1996 . A comparison of approaches for maximizing business payoff of prediction models",In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, Oregon, USA, pp. 195–201.

Masethe, D.,H., Masethe, M. 2014. Prediction of heart disease using classification algorithms. *proceedings of the world congress on engineering and computer science vol II ,*San Francisco, USA.

Mduma, N., Kalegele, K. and Machuve, D. 2019. A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction. *Data Science Journal*, 18(1), p.14.

Metha, M., Agrawal, R. and Riassnen, J. 1996. SLIQ: A fast scalable classifier for data mining," Extending Database Technology, pp. 18-32.

Miao, J., Niu, L. 2016 . A Survey on Feature Selection. *Information Technology and Quantitative Management (ITQM 2016), Procedia Computer Science 91 (2016) 919 – 926 , Elsevier.*

Namburu, S.M., Tu, H., Luo, J., Pattipati, K.R. 2005. Experiments on Supervised Learning Algorithms for Text Categorization . IEEE Aerospace Conference.

Ozcift A. 2011. SVM Feature Selection Based Rotation Forest Ensemble Classifiers to Improve Computer-Aided Diagnosis of Parkinson Disease. Journal of Medical Systems.

Pang-Ningtan, Kumar, V., Steinbach, M., 2008.Introduction to Data mining" Pearson.
Pao, Y., & Sobajic, D. J. 1992.The combined use of unsupervised and supervised learning for dynamic security assessment" Transactions on Power Systems, 7(2), 878-884.

Patil, B., S. 2009. Extraction of significant patterns from heart disease warehouses for heart attack prediction. *International journal of computer science and network security, vol.9*.

Pratiwi, A. I. , Adiwijaya,(2018), " On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis ", Applied Computational Intelligence and Soft Computing, 1687-9724, Hindawi.

Qinbao S, Jingjie N, and Wang G. 2013.A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. IEEE transactions on knowledge and data engineering, vol. 25, no. 1.
Queen, M. J., 1967. Some methods for classification and analysis of multivariate observations. Proc. of the 5th Berkeley Symp. Math. Statist, pp. 281-297.

Quinlan, J., R. 1986. Induction of decision trees. *Machine Learning, vol. 1, no. 1, pp. 81–106.*

Ramana, V.N., Murty, and Prasadbabu M.S. 2017. A Critical Study of Classification Algorithms for Lung CancerDisease Detection and Diagnosis. *International Journal of Computational Intelligence Research*, Research India Publications ISSN 0973-1873 Volume 13, pp. 1041-1048.

Rangra, K., and Bansal, K. L. 2014. Comparative Study of Data Mining Tools. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6.

Rao, R., 2003. Data Mining and Clustering Techniques. *DRTC Workshop on Semantic Web 8th – 10th December*, Bangalore.

Rastogi, R., and Shim, K. 1998.Public: A decision tree classifier that integrates building and pruning," Proc. of the 24th International Conference on Very Large Data Bases, pp. 404-415.

Rathee A, Mathur R.P. 2013. Survey on Decision Tree Classification algorithm for the Evaluation of Student Performance. *International Journal of Computers & Technology.*

Riquelme, J. C., Ruiz, R., Rodriguez, D., and Moreno, J. 2008. Finding defective modules from highly unbalanced datasets, Act as de Los Talleres de.

Rout, N., Mishra, D., Mallick, M. K., and Reddy M.S. 2018. Handling Imbalanced Data: A Survey. International Proceedings on Advances in Soft Computing, Intelligent Systems, and Applications, Advances in Intelligent Systems and Computing 628, Springer Nature Singapore.

RuiXu 2005. Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks, Vol. 16.*

Saeys, Y., Inza, A., and Larran, P. 2007. A review of feature selection techniques in bioinformatics. *BIOINFORMATICS REVIEW Vol. 23 no. 19 pages 2507–2517*.

Sarvestani, S. A, Safavi, A. A., Parandeh, N.M., Mansoor, S. 2010. Predicting breast cancer survivability using data mining techniques.

Sethunya, R., Joseph, L. 2016. Data Mining Algorithms: An Overview. *International journal of computers and technology council for Innovative Research* Volume 15 April.

Shah, N. K. ,Paul J.,Greenville, G. 1989. Calculating Mahalanobis distances using principal component analysis trends in analytical chemistry", vol. 8.

Sharma, D., Sharma, A., Mansotra V. 2017. A Literature Survey on Data Mining Techniques to Predict Lifestyle. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* Volume-5 ISSN: 2321-9653.

Shaza M. Elrahman, A., and Abraham, A. 2013. A Review of Class Imbalance Problem. Journal of Network and Innovative Computing ISSN 2160-2174, Volume 1 pp. 332-340.

Sheikhpour, R., Gharaghani, M., Zare, M. A., Chahookia .2017. A Survey on semi-supervised feature selection methods.Pattern Recognition Volume 64, April 2017, Pages 141-158.

Shekhar R et al. 2007 .K-Means+ID3: A Novel Method for Supervised Anomaly Detection by Cascading K-means clustering and ID3 Decision Tree Learning Methods. IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 3.

Sherry, Chen, Y. and Liu X. 2004. The Contribution of Data Mining in Information Science.*Journal of Information Science, © CILIP.*

Singh, Y. and Chauhan, A. S. 2009. Neural Networks in Data Mining.Journal of Theoretical and Applied

Information Technology.

Smitha. T, Sundaram, V. 2012. Comparative Study of Data Mining Algorithms for High Dimensional Data Analysis International Journal of Advances in Engineering & Technology,.IJAET ISSN: 2231-1963.

Smuc, T., Gamberger, D., and Krstacic, G. 2001. Combining unsupervised and supervised machine learning in analysis of the CHD patient database. The American Invitational Mathematics Examination, 109–112.

Sumana, B.V., and Santhanam, T. 2014.Prediction of diseases by Cascading Clustering and Classification.International Conference on Advances in Electronics, Computers, and Communications (ICAECC) IEEE.

Sumana, B.V., and Santhanam, T. 2016. Predictionction of imbalanced data using a Cluster-based Approach.Asian Journal of Information technology 15(16):3022-3042, ISSN: 1682-3915 @ medwell journals.

Sumana, B.V., Santhanam, T. 2010.An Empirical Comparison of Ensemble and Hybrid Classification.Proc. of Int. Conf. on Recent Trends in Signal Processing, Image Processing, and VLSI.

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., Zhou, Y. 2015. A novel ensemble method for classifying imbalanced data. Pattern Recognition 48, 1623–1637.

Swarndeep S. J, Pandya, S. 2016. An Overview of Partitioning Algorithms in Clustering Techniques. International Journal of Advanced Research in Computer Engineering & Technology.

Thiago F. C., Hruschka, E. R. 2011.Towards improving cluster-based feature selection with a simplified silhouette filter" Information Sciences 181 (2011) 3766–3782.

Tong, S., and Koller, D. 2001. Support Vector Machine Active Learning with Applications to Text Classification. Journal of Machine Learning Research 45-66.

Tutorial on Support Vector Machine (SVM) Vikramaditya Jakkula, School of EECS, Washington State University, Pullman 99164.

UCI machine learning repository, available at http: //archive.ics.uci.edu/ml.

Urbanowicza, R. J., Meekerb, M., Cavaa, W. L., Olsona, R. S., Moorea, J.H. 2018. Relief-based feature selection: Introduction and review.Journal of Biomedical Informatics, Volume 85, Pages 189-203, ISSN 1532-0464.

Verdu S. 1998.Fifty years of Shannon's theory.IEEE Transactions on Information Theory. 2057-78.

Verma, A., Kaur, I., Singh I., 2016. Comparative Analysis of Data Mining Tools and Techniques for Information Retrieval.*Indian Journal of Science and Technology*, *Vol 9(11), DOI: 10.17485/ijst/2016/v9i11/81658.*

Vijayalakshmi, D., Thilagavathi, K., Breault, J. 2012. Data Mining Diabetic Databases: are rough sets a useful addition? An approach for prediction of diabetic disease by using the b-coloring technique in clustering analysis. *International Journal of applied mathematical research* science publishing corporation, pp.520-530.

Wang, Q., Luo, Z., Huang, J., Feng, Y. and Zhong Liu, Z. 2017. A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM. *Computational Intelligence and Neuroscience Hindawi, Article ID 1827016.*

Wilkinson, L., 1992.Tree-Structured Data Analysis: AID, CHAID, and CART. Paper presented at the Sun Valley, ID, Sawtooth/SYSTAT Joint Software Conference.

Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S.J. 1999. Weka: Practical machine learning tools and techniques with Java implementations. (Working paper 99/11). Hamilton, New Zealand: the University of Waikato, Department of Computer Science.

Wu, X. et al. 2007. Top 10 algorithms in data mining.*Springer-Verlag London Limited*.

Wu, X. et. al 2008.Top 10 algorithms in data mining. Knowledge and Information system, pp1-37.

Xie W., Liang G, Dong Z, Tan B., Zhang B. 2019. An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data. Mathematical Problems in Engineering.

Xu, D., Tian, Y. 2015. A Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci., Springer-Verlag Berlin Heidelberg*, pp. 165–193.

Li, Y., Hung, E., Chung, K., Huang, J. 2008. Building A Decision Cluster Classification Model for High Dimensional Data by A Variable Weighting k-Means Method. Proc. of the Twenty-First Australasian Joint Conference on Artificial Intelligence, pp. 337-347.

Yang, Q. and Xindong W. 10 challenging problems in data mining research. international journal of information technology & decision making ISSN (print): 0219-6220 | ISSN (online): 1793-6845.

Yen, S., Lee, Y. 2009. Cluster-based under-sampling approaches for imbalanced data distributions.*Expert Systems with Applications 36, 5718–5727*.

Yong, Z., Youwen, L., and Shixiong, X. 2009. An Improved KNN Text Classification Algorithm Based on clustering. Journal of Computers, vol. 4, no. 3.

Yvan S, Inza A and Larran, P. 2007. A review of feature selection techniques in bioinformatics. BIOINFORMATICS REVIEW Vol. 23 no. 19 2007, pages 2507–2517.

Zehra, A. 2014.A comparative study on the pre-processing and mining of Pima Indian Diabetes Dataset, ICSEC, The International Computer Science and Engineering Conference (ICSEC) 1-10.

Zeng, H. et al 2003. CBC: Clustering Based Text Classification Requiring Minimal Labeled.Proceedings of the Third IEEE International Conference on Data Mining (ICDM'03) IEEE.

Zhou and Liu, X.Y. 2006.Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE transactions on Knowledge and Data Engineering, Vol 18, PP 63-77,2006.

Zhou, L., Lai, K. K.2009. Benchmarking binary classification models on data sets with different degrees of imbalance. Comput.Sci China 3(2):205-216.

Mrs. Subodhini Gupta
Department of Computer Science and Engineering
Faculty of Technology and Engineering
The Maharaja Sayajirao University of Baroda
Vadodara 390001.