AN ENHANCED RESAMPLING TECHNIQUE FOR IMBALANCED DATA SETS

MAISARAH BINTI ZORKEFLEE

MASTER OF SCIENCE (INFORMATION TECHNOLOGY)
UNIVERSITI UTARA MALAYSIA
2015

Permission to Use

In presenting this thesis in fulfilment of the requirements for a postgraduate degree from Universiti Utara Malaysia, I agree that the Universiti Library may make it freely available for inspection. I further agree that permission for the copying of this thesis in any manner, in whole or in part, for scholarly purpose may be granted by my supervisor(s) or, in their absence, by the Dean of Awang Had Salleh Graduate School of Arts and Sciences. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Awang Had Salleh Graduate School of Arts and Sciences
UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 UUM Sintok

Abstrak

Set data adalah tidak seimbang apabila sampel data yang terdapat pada satu kelas (kelas majoriti) melebihi kelas selain daripadanya (kelas minoriti). Masalah utama berkaitan dengan data binari tidak seimbang ialah kecenderungan pengelas untuk mengabaikan kelas minoriti. Beberapa teknik persampelan semula seperti pensampelan bawah, pensampelan atas dan gabungan kedua-duanya telah banyak digunakan. Walau bagaimanapun, teknik pensampelan bawah dan pensampelan atas tersebut masih mempunyai kekurangan seperti pembuangan dan penambahan data yang berguna yang menyebabkan masalah ketepatan pengelasan data. Oleh itu, kajian ini bertujuan untuk meningkatkan metrik klasifikasi dengan menambah baik teknik pensampelan bawah dan menggabungkannya dengan teknik pensampelan atas yang telah wujud. Untuk mencapai objektif tersebut, teknik Pensampelan Bawah Berdasarkan Jarak Kabur (FDUS) dicadangkan. Anggaran entropi digunakan untuk menghasilkan ambang kabur untuk mengelaskan sampel di dalam kelas minoriti dengan kelas majoriti kepada fungsi keahlian. FDUS kemudian digabungkan dengan Teknik Pensampelan Atas Minoriti Sintetik (SMOTE) dikenali sebagai FDUS+SMOTE, dilakukan di dalam urutan sehingga data yang seimbang dihasilkan. Kedua-dua teknik, FDUS and FDUS+SMOTE dibandingkan dengan empat teknik yang lain berdasarkan ketepatan klasifikasi, F-ukuran dan G-purata. Berdasarkan keputusan, FDUS mencapai ketepatan klasifikasi F-ukuran dan G-purata yang lebih bagus apabila dibandingkan dengan teknik lain dengan purata masing-masing 80.57%, 0.85 dan 0.78. Ini menunjukkan logik kabur apabila digabungkan dengan teknik Pensampelan Bawah Berdasarkan Jarak mampu mengurangkan penyingkiran data yang berguna. Tambahan, penemuan menunjukkan FDUS+SMOTE menghasilkan prestasi yang lebih baik berbanding gabungan teknik SMOTE dan Pautan Tomek, dan SMOTE dan Penyuntingan Jiran Terdekat pada data penanda aras. FDUS+SMOTE telah mengurangkan pembuangan data yang berguna dari kelas majoriti dan mengelakkan terlebih-padanan. Secara purata, FDUS dan FDUS+SMOTE mampu mengimbangkan data kategorik, integer dan nyata serta membaiki prestasi klasifikasi binari. Selain itu, teknik tersebut menghasilkan prestasi yang baik pada data yang mempunyai saiz rekod kecil yang mempunyai sampel di dalam lingkungan kira-kira 100 ke 800.

Kata kunci: Data tidak seimbang, Teknik persampelan semula, Teknik pensampelan bawah, Teknik pensampelan atas, Logik kabur

Abstract

A data set is considered imbalanced if the distribution of instances in one class (majority class) outnumbers the other class (minority class). The main problem related to binary imbalanced data sets is classifiers tend to ignore the minority class. Numerous resampling techniques such as undersampling, oversampling, and a combination of both techniques have been widely used. However, the undersampling and oversampling techniques suffer from elimination and addition of relevant data which may lead to poor classification results. Hence, this study aims to increase classification metrics by enhancing the undersampling technique and combining it with an existing oversampling technique. To achieve this objective, a Fuzzy Distance-based Undersampling (FDUS) is proposed. Entropy estimation is used to produce fuzzy thresholds to categorise the instances in majority and minority class into membership functions. FDUS is then combined with the Synthetic Minority Oversampling TEchnique (SMOTE) known as FDUS+SMOTE, which is executed in sequence until a balanced data set is achieved. FDUS and FDUS+SMOTE are compared with four techniques based on classification accuracy, F-measure and G-mean. From the results, FDUS achieved better classification accuracy, F-measure and G-mean, compared to the other techniques with an average of 80.57%, 0.85 and 0.78, respectively. This showed that fuzzy logic when incorporated with Distance-based Undersampling technique was able to reduce the elimination of relevant data. Further, the findings showed that FDUS+SMOTE performed better than combination of SMOTE and Tomek Links, and SMOTE and Edited Nearest Neighbour on benchmark data sets. FDUS+SMOTE has minimised the removal of relevant data from the majority class and avoid overfitting. On average, FDUS and FDUS+SMOTE were able to balance categorical, integer and real data sets and enhanced the performance of binary classification. Furthermore, the techniques performed well on small record size data sets that have of instances in the range of approximately 100 to 800.

Keywords: Imbalanced data, Resampling technique, Undersampling technique, Oversampling technique, Fuzzy logic

Acknowledgement

All praise to Allah who gave me patience and strength to complete this study.

I would like to take this opportunity to express my gratitude to my main supervisor, Miss Aniza Mohamed Din for her advice and encouragement. A special thanks to my co-supervisor Prof. Dr. Ku Ruhana Ku Mahamud for her guidance throughout the completion of my study. I would also like to thank the appointed examiners for their valuable critiques to improve my thesis.

To my parents, Zorkeflee Abu Hasan and Badariah Mohd Yusoff, thank you for your love and prayers. To my husband, Zariq Zaquan Razani, thank you for your endless support and encouragement. To my siblings, Zuhaili, Zulhilmi and Mahirah, thank you for your inspirations and words of wisdom to boost up my spirit. I really appreciate all of you.

Last but not least, I wish to thank all my friends especially my labmates, who continuously help, support and motivate me during this journey.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AUC	Area Under ROC Curve
BRACID	Bottom-up induction of Rules and Cases for Imbalanced Data
CNN	Condensed Nearest Neighbour
DID	Department of Irrigation and Drainage
DUS	Distance-based Under-Sampling
ENN	Edited Nearest Neighbor
EUS	Evolutionary Undersampling
FDUS	Fuzzy Distance-based Undersampling
FDUS+SMOTE	FDUS and SMOTE
FN	False Negative
FP	False Positive
G-mean	Geometric mean
ISMOTE	Improved SMOTE
ISMOTE+DUS	ISMOTE and DUS
k-NN	k-Nearest Neighbour
m	meter
mm	millimetre
MSMOTE	Modified SMOTE
NCL	Neighbourhood Cleaning Rule
OSS	One-Sided Selection
RNN	Reduced Nearest Neighbour
ROC	Receiver Operating Characteristics
ROS	Random Over-Sampling
RUS	Random Under-Sampling
SMOTE	Synthetic Minority Over-sampling TEchnique
SMOTE+ENN	SMOTE and ENN
SMOTE+TL	SMOTE and TL
SVM	Support Vector Machine
TL	Tomek Links
TN	True Negative
TP	True Positive

CHAPTER ONE
INTRODUCTION

Data is a set of values of qualitative and quantitative variables in order to deliver information. Often, the distribution of the data sets are imbalanced. This chapter provides some background about imbalanced data sets and the problem related to them. The research objectives, research scope and significance of study are also stated in this chapter.

1.1 Background

Imbalanced data sets occur when the number of samples in one class is low as compared to other classes (Barua, Islam, Yao, & Murase, 2014). In binary classification, the class that contain less instances is known as minority class, and the other class is known as majority class. Examples of imbalanced data sets are flood events (Wang, Chen, & Small, 2013), medical data sets (Dubey, Zhou, Wang, Thompson & Ye, 2014), intrusion detection data sets (Chairi, Alaoui, & Lyhyaoui, 2012), credit card fraud detection (Padmaja, Dhulipalla, Krishna, Bapi, & Laha, 2007), and oil spill identification (Brekke & Solberg, 2005). The issue that is commonly related to imbalanced data is poor classification performance due to the tendency of classifiers to ignore data samples that belong to the minority class (Lin & Chen, 2012; Mangai, Samanta, Das, & Chowdhury, 2010; Mi, 2013). For example, when imbalanced data is classified using Support Vector Machine (SVM), the decision boundary obtained is biased towards the minority class resulting to misclassification (Liu, Yu, Huang, & An, 2011; Bennett & Bredensteiner, 2000). This bias will reduce the performance of SVM with respect to the minority class (Batuwita & Palade, 2013).

The contents of the thesis is for internal user only

REFERENCES

Alejo, R., Garcia, V., Sotoca, J. M., Mollineda, R. A., & Sanchez, J. S. (2007). Improving the performance of the RBF neural networks trained with imbalanced samples. *Computational and Ambient Intelligence*, *4507*, 162–169.

Anand, A., Pugalenthi, G., Fogel, G. B., & Suganthan, P. N. (2010). An approach for classification of highly imbalanced data using weighting and undersampling. *Amino acids*, *39*(5), 1385-1391.

Aziz, A. M. (2009, August). Effects of fuzzy membership function shapes on clustering performance in multisensor-multitarget data fusion systems. In *Fuzzy Systems, 2009. FUZZ-IEEE 2009* (pp. 1839-1844).

Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.

Barua, S., Islam, M., Yao, X., & Murase, K. (2014). MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. *Knowledge and Data Engineering, IEEE Transactions on*, *26*(2), 405-425.

Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003, December). Balancing training data for automated annotation of keywords: a case study. In *WOB*, 10-18.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20-29.

Batuwita, R., & Palade, V. (2013). Class imbalance learning methods for support vector machines. *Imbalanced learning: Foundations, algorithms, and applications*, 83-99

Bedient, P. B., Huber, W. C., & Vieux, B. E. (2008). Hydrology and floodplain analysis fourth edition. Prentice Hall.

Bekkar, M., & Alitouche, T. A. (2013). Imbalanced data learning approaches. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *3*(4), 15–33.

Bennett, K. P., & Bredensteiner, E. J. (2000, June). Duality and geometry in SVM classifiers. In *ICML* (pp. 57-64).

Brekke, C., & Solberg, A. H. S. (2005). Oil spill detection by satellite remote sensing. *Remote Sensing of Environment*, *95*(1), 1–13.

Chairi, I., Alaoui, S., & Lyhyaoui, A. (2012). Learning from imbalanced data using methods of sample selection. In *Multimedia Computing and Systems (ICMCS)*, 254-257. IEEE.

Carvajal, K., Chacon, M., Mery, D., & Acuna, G. (2004) Neural network method for failure detection with skewed class distribution. *INSIGHT, Journal of the British Institute of Non-Destructive Testing*, *46*(7), 399–402.

Chawla, N. V. (2010). Data mining for imbalanced data sets: an overview. In *Data Mining and Knowledge Discovery Handbook*, 875-886. Springer US.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE : Synthetic Minority Over-sampling TEchnique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1-6.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. (2003). SMOTEBoost : Improving prediction of the minority class in boosting. *Proceedings Principles Knowledge Discovery Databases* (pp. 107–119).

Chiang, H. S., Shih, D. H., Lin, B., & Shih, M. H. (2014). An APN model for arrhythmic beat classification. *Bioinformatics*, *30*(12), 1739-1746.

Christensen, R. (1980). Entropy Minimax Sourcebook. Vol. 1–4, Entropy Ltd., Lincoln, MA.

Del Gaudio, R., Batista, G., & Branco, A. (2014). Coping with highly imbalanced datasets: A case study with definition extraction in a multilingual setting. *Natural Language Engineering*, *20*(03), 327-359.

Diamantini, C., & Potena, D. (2009). Bayes vector quantizer for class-imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, *21*(5), 638–651.

Ding, Z. (2011). Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics, in Computer Science Department, Georgia State University.

Dubey, R., Zhou, J., Wang, Y., Thompson, P. M., & Ye, J. (2014). Analysis of sampling techniques for imbalanced data : An n = 648 ADNI study. *NeuroImage*, *87*, 220–241.

Ertekin, S., Huang, J., Bottou, L., & Giles, L. (2007). Learning on the border: active learning in imbalanced data classification. In *Proceedings of the sixteenth ACM* *conference on Conference on information and knowledge management* (pp. 127-136). ACM.

Fernandez, A., Del Jesus, M. J., & Herrera, F. (2009). Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets. *International Journal of Approximate Reasoning*, *50*(3), 561–577.

Fernandez, A., Lopez, V., Galar, M., Del Jesus, M. J., & Herrera, F. (2013). Analysing the classification of imbalanced data-sets with multiple classes: binarization techniques and ad-hoc approaches. *Knowledge-Based Systems*, *42*, 97-110.

Fitkov-Norris, E., & Folorunso, S. O. (2013). Impact of sampling on neural network classification performance in the context of repeat movie viewing. *EANN 2013, Part I, CCIS 383*, 213–222.

Folorunso, S. O. & Adeyemo, A. B. (2012). Theoretical comparison of undersampling techniques against their underlying data reduction techniques. *EIECON2012*, 92-97.

Fu, X., Wang, L., Chua, K. S., & Chu, F. (2002). Training RBF neural networks on unbalanced data. *Proceedings of the 9th International Conference on Neural Information Processing*, *2*, 1016–1020.

Galar, M., Fernandez, A., Barrenechea, E., & Herrera, F. (2013). EUSBoost: enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling. *Pattern Recognition*, *46*(12), 3460-3471.

Ganesh, M. (2006). Introduction to fuzzy sets and fuzzy logic. India, ND: Prentice-Hall of India Private Limited.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced data sets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42–47.

Garcia, V., Mollineda, R. A., & Sanchez, J. S. (2008). On the k-NN performance in a challenging scenario of imbalance and overlapping. *Pattern Analysis and Applications*, *11*(3-4), 269–280.

Garcia, V., Sanchez, J. S., Mollineda, R. A., Alejo, R., & Sotoca, J. (2007). The class imbalance problem in pattern classification and learning. *Congreso Espanol de Informatica*, (pp. 284–291).

Gates, G. W. (1971). The reduced nearest neighbor rule. *IEEE Trans Information Theory*, *18*(3), 431–433.

Goel, G., Maguire, L., Li, Y., & McLoone, S. (2013). Evaluation of sampling methods for learning from imbalanced data. In *Intelligent Computing Theories (*pp. 92-401). Springer Berlin Heidelberg.

Gu, Q., Cai, Z., & Zhu, L. (2009). Classification of imbalanced data sets by using the hybrid re-sampling algorithm based on isomap. In *Advances in Computation and Intelligence* (pp. 287-296). Springer Berlin Heidelberg.

Hart, P. E. (1968). The condensed nearest neighbour rule. *IEEE Transactions on Information Theory*, 515–516.

He, G., Han, H., & Wang, W. (2005). An over-sampling expert system for learning from imbalanced data sets. *Neural Networks and Brain, 2005. ICNN&B '05,* (pp. 537–541). Beijing.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Hu, S., Liang, Y., Ma, L., & He, Y. (2009). MSMOTE: improving classification performance when training data is imbalanced. *2009 Second International Workshop on Computer Science and Engineering* (pp. 13–17).

Hu, X., Lin, T. Y., & Han, J. (2004). A new rough sets model based on database systems. *Fundamenta Informaticae*, *59*(2), 135-152.

Jeatrakul, P., & Wong, K. W. (2012). Enhancing classification performance of multi-class imbalanced data using the OAA-DB algorithm. *The 2012 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE.

Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm. In *Neural Information Processing. Models and Applications* (pp. 152-159). Springer Berlin Heidelberg.

Jiang, W., Deng, L., Chen, L., Wu, J., & Li, J. (2009). Risk assessment and validation of flood disaster based on fuzzy mathematics. *Progress in Natural Science*, *19*(10), 1419–1425.

Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*(1), 18–36.

Kanagavalli, V. R., & Raja, K. (2011). Detecting and resolving spatial ambiguity in text using named entity extraction and self-learning fuzzy logic techniques.

Kim, D.-S., Baek, Y.-M., & Kim, W.-Y. (2013). Reducing overfitting of AdaBoost by clustering-based pruning of hard instances. *Proceedings of the 7th*

International Conference on Ubiquitous Information Management and Communication - ICUIMC ’13 (pp. 1–3).

Kim, M. (2013). Geometric mean based boosting algorithm to resolve data imbalance problem. *The Fifth International Conference on Advances Databases, Knowledge and Data Applications* (pp. 15–20).

Kubat, M., & Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided selection. *Proceedings of the fourteenth conference on machine learning* (pp. 179–186).

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution, 63-66. Springer Berlin Heidelberg.

Lee, C. Y., Yang, M. R., Chang, L. Y., & Lee, Z. J. (2010). A hybrid algorithm applied to classify unbalanced data. In *Networked Computing and Advanced Information Management (NCM) (*pp. 618-621). IEEE.

Lee, C., & Lee, Z. (2012). A novel algorithm applied to classify unbalanced data. *Applied Soft Computing Journal*, *12*(8), 2481–2485.

Li, D.-C., Liu, C.-W., & Hu, S. C. (2010). A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine*, *40*(5), 509-518.

Li, H., Zou, P., Wang, X., & Xia, R. (2013). A new combination sampling method for imbalanced data. In *Proceedings of 2013 Chinese Intelligent Automation Conference* (pp. 547-554). Springer Berlin Heidelberg.

Lin, W. J., & Chen, J. J. (2012). Class-imbalanced classifiers for high-dimensional data. *Briefings in bioinformatics*, *14*(1), 13-26.

Liu, X., Wu, J., & Zhou, Z. (2009). Exploratory undersampling for class-imbalance learning, *39*(2), 539–550.

Liu, W., Chawla, S., Cieslak, D. A., & Chawla, N. V. (2010). A Robust Decision Tree Algorithm for Imbalanced Data Sets. In *SDM*, *10*, 766-777.

Liu, Y., Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing & Management*, *47*(4), 617-631.

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,* 113–141.

Luengo, J., Fernandez, A., Garcia, S., & Herrera, F. (2011). Addressing data complexity for imbalanced data sets : analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*, *15*, 1909–1936.

Mahdizadeh, M., & Eftekhari, M. (2013). Designing fuzzy imbalanced classifier based on the subtractive clustering and genetic programming. *Iranian Conference on Fuzzy Systems (IFSC)* (pp. 8–13).

Mangai, U. G., Samanta, S., Das, S., & Chowdhury, P. R. (2010). A survey of decision fusion and feature fusion strategies for pattern classification. *IETE Technical Review*, *27*(4), 293–307.

Mann, P. S. (2012). *Introductory Statistics* (8th ed.).Wiley Global Education.

Mi, Y. (2013). Imbalanced classification based on active learning SMOTE. *Research Journal on Applied Sciences, Engineering and Technology*, *5*(3), 944–949.

Mirza, B., Lin, Z., & Toh, K. A. (2013). Weighted online sequential extreme learning machine for class imbalance learning. *Neural Processing Letters*, 1-22.

Naganjaneyulu, S., & Kuppa, M. R. (2012). A novel framework for class imbalance learning using intelligent under-sampling. *Progress in Artificial Intelligence*, *2*(1), 73–84.

Napierala, K., & Stefanowski, J. (2012). BRACID: A comprehensive approach to learning rules from imbalanced data. *Journal of Intelligent Information Systems*, *39*(2), 335–373.

Nguyen, G. H., Bouzerdoum, A., & Phung, S. L. (2009). Learning pattern classification tasks with imbalanced data sets. In *P. Yin (Eds), Pattern Recognition* (pp. 193-208). Vukovar, Croatia: In-Teh.

Orriols-Puig & Bernadó-Mansilla (2009). Evolutionary rule-based systems for imbalanced data sets. *Soft Computing*, *13*(3), 213-225.

Ou-Yang, C., Rieza, M., Wang, H.-C., Juan, Y.-C., & Huang, C.-T. (2013). Applying a hybrid data preprocessing methods in stroke prediction. In Y.-K. Lin, Y.-C. Tsao, & S.-W. Lin (Eds.), *Proceedings of the Institute of Industrial Engineers Asian Conference 2013* (pp. 1441–1449). Singapore: Springer Singapore.

Padmaja, T. M., Dhulipalla, N., Krishna, P. R., Bapi, R. S., & Laha, A. (2007). An unbalanced data classification model using hybrid sampling technique for fraud detection. In *Pattern Recognition and Machine Intelligence* (pp. 341-348). Springer Berlin Heidelberg.

Phung, S. L., Bouzerdoum, A., & Nguyen, G. H. (2009). Learning pattern classification tasks with imbalanced data sets. *Pattern Recognition*, 93–208.

Rokach, L. (2009). Ensemble-based classifiers. *Artificial Intelligence Review*, *33*(1-2), 1–39.

Ross, T. J. (2010). Development of membership functions. *Fuzzy Logic with Engineering Applications, Third Edition*, 174-210.

Sang, G., Gao, L., & Liu, Z. (2013). A bias-ensemble learning algorithm for imbalanced data processing imbalanced data-sets classification methods. *Journal of Computational Information Systems*, *9*(5), 2025–2032.

Segretier, W., Clergue, M., Collard, M., & Izquierdo, L. (2012). An evolutionary data mining approach on hydrological data with classifier juries. *2012 IEEE Congress on Evolutionary Computation*, 1–8.

Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). RUSBoost: a hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *40*(1), 185–197.

Shen, Q., & Jiang, Y. (2010). Fuzzy sets, rough sets and vague sets. *3rd International Conference onAdvanced Computer Theory and Engineering* (pp. 461–465).

Shivalkar, P. S., & Tripathy, B. K. (2015). Rough Set Based Green Cloud Computing in Emerging Markets.

Singpurwalla, N. D., & Booker, J. M. (2004). Membership functions and probability measures of fuzzy sets. Journal of the American Statistical Association, 99(467), 867-877.

Sivanandam, S. N., Sumathi, S., & Deepa, S. N. (2007). Introduction to fuzzy logic using matlab. Berlin, Heidelberg: Springer Berlin Heidelberg.

Soler, V. & Prim, M. (2009). Extracting a fuzzy system by using genetic algorithms for imbalanced data sets classification: application on down syndrome detection. In D. A. Zighed, S. Tsumoto, Z. W. Ras, & H. Hacid. (Eds.), *Mining Complex Data* (pp. 23-39). Springer Berlin Heidelberg.

Sun, Y., Kamel, M. S., Wong, A. K. C., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, *40*(12), 3358–3378.

Sun, Y., Robinson, M., Adams, R., Boekhorst, R., Rust, A. G., & Davey, N. (2006). Using sampling methods to improve binding site predictions. *Procs of the 14th European Symposium on Artificial Neural Networks, ESANN 2006* (pp. 533–538).

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data : a review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). Ensemble methods. *Introduction to data mining*, 276–293. United States of America: Pearson Education.

Tang, Y., Zhang, Y. Q., Chawla, N. V., & Krasser, S. (2009). SVMs modeling for highly imbalanced classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *39*(1), 281-288.

Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. *IEEE Transaction on System, Man, and Cybernetics*, *6*(6), 448–452.

Verbiest, N., Ramentol, E., Cornelis, C., & Herrera, F. (2012). Improving SMOTE with fuzzy rough prototype selection to detect noise in imbalanced, 169–178.

Visa, S., & Ralescu, A. (2003). Learning imbalanced and overlapping classes using fuzzy sets. *Workshop on Learning from Imbalanced Datasets II (ICML '03)* (pp. 91–104).

Visa, S., & Ralescu, A. (2005, April). Issues in mining imbalanced data sets-a review paper. In *Proceedings of The Sixteen Midwest Artificial Intelligence And Cognitive Science Conference* (pp. 67-73).

Wang, D., Chen, P., & Small, D. L. (2013). Towards long-lead forecasting of extreme flood events : a data mining framework for precipitation cluster precursors identification, 1285–1293.

Wang, S., & Yao, X. (2009, March). Diversity analysis on imbalanced data sets by using ensemble models. In *Computational Intelligence and Data Mining, 2009. CIDM'09* (pp. 324-331). IEEE.

Wang, X. J., Zhao, R. H., & Hao, Y. W. (2011). Flood control operations based on the theory of variable fuzzy sets. *Water Resources Management*, *25*(3), 777–792.

Wang, S., & Yao, X. (2013). Relationships between diversity of classification ensembles and single-class performance measures. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), 206–219.

Weiss, G. M., & Provost, F. (2003). Learning when training data are costly: the effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 315-354.

Whitley, E., & Ball, J. (2001). Statistics review 1: presenting and summarising data. *Critical Care*, *6*(1), 66.

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, *2*(3), 408–421.

Wong, G. Y., Leung, F. H., & Ling, S. H. (2014, July). An under-sampling method based on fuzzy logic for large imbalanced dataset. In *Fuzzy Systems (FUZZ-IEEE)* (pp. 1248-1252). IEEE.

Yang, H., Fong, S., Wong, R., & Sun, G. (2013). Optimizing classification decision trees by using weighted naive bayes predictors to reduce the imbalanced class problem in wireless sensor network. *International Journal of Distributed Sensor Networks, 2013*, 1–16.

Yang, Z., & Gao, D. (2013). Classification for imbalanced and overlapping classes using outlier detection and sampling techniques. *Applied Mathematics & Information Sciences, 7*(1L), 375–381.

Zadeh, L. A. (1980). Fuzzy sets versus probability. *Proceedings of the IEEE, 68*(3), 421.

Zhang, D., Liu, W., Gong, X., & Jin, H. (2011). A novel improved smote resampling algorithm based on fractal. *Journal of Computational Information Systems*, *6*, 2204–2211.

Zhang, I., & Mani, I. (2003). kNN approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of Workshop on Learning from Imbalanced Datasets*.

Zhang, Y., & Wang, D. (2013). A cost-sensitive ensemble method for class-imbalanced data sets. *Abstract and Applied Analysis*, *2013*, 1–6.

Zhong, W., Raahemi, B., & Liu, J. (2009). Learning on class imbalanced data to classify peer-to-peer applications in IP traffic using resampling techniques. In *Neural Networks, 2009. IJCNN 2009*, (pp. 3548-3554). IEEE.