Imbalanced Classification for Business Analytics

Talayeh Razzaghi

Department of industrial Engineering and Management Systems

University of Central Florida

4000 Central Florida Boulevard

Orlando, Florida 32816-2993

USA

email: rtalayeh@knights.ucf.edu

Andrea Otero

Department of industrial Engineering and Management Systems

University of Central Florida

4000 Central Florida Boulevard

Orlando, Florida 32816-2993

USA

email: andrea\_otero\_2012@knights.ucf.edu

Petros Xanthopoulos\*

Department of industrial Engineering and Management Systems

University of Central Florida

4000 Central Florida Boulevard

Orlando, Florida 32816-2993

USA

voice: +1 407-823-5218

fax: +1 407-823-3413

email: petrosx@ucf.edu

(\* Corresponding author)

Imbalanced Classification for Business Analytics

**Talayeh Razzaghi**

*University of Central Florida, USA*

**Andrea Otero**

*University of Central Florida, USA*

**Petros Xanthopoulos**

*University of Central Florida, USA*

**INTRODUCTION**

In pattern recognition, classification is a crucial task for automated data driven knowledge discovery. The objective of classification is to separate a set of data into classes or sub-categories and then to identify the classes that a new observation belongs to according to a training set of data. The mathematical model trained by a classification algorithm is termed *classifier*. When the class size of given examples is not equal for all classes, the classification problem is known as imbalanced ([Japkowicz, 2000](#_ENREF_26)). For instance, in a cancer diagnostic problem the main objective is to identify individuals stricken with cancer and such events are relatively rare compared to normal cases. Imbalanced classification problems are also known as *skewed class distribution problems* or as *small/ rare class learning problems* ([He & Garcia, 2009](#_ENREF_20); [Lemnaru & Potolea, 2012](#_ENREF_31); [Sun, Wong, & Mohamed, 2009](#_ENREF_45)). In binary classification, the class with fewer examples is known as the *minority class* and the other class as the *majority class*. In many applications (e.g. fraud detection, computer intrusion detection, oil spill detection, defect product detection), detection of minority class examples is more important than the majority class. Therefore, there is a need for efficient classification algorithms to address such problems. A preferred classification algorithm is the one that yields higher identification rate on rare events especially for applications where their misclassification yields to high losses. For instance in automated credit card fraud detection, a fraud event misclassification might result in high monetary losses for the credit card vendor. On the other side misclassification of non-fraudulent events will worsen the customer satisfaction experience.

The emerging nature of imbalanced classification problems has led to the development of modified algorithms and new performance metrics. Standard performance measures such as classification accuracy are not appropriate when the data is imbalanced ([N. V. Chawla, 2010](#_ENREF_10); [He & Garcia, 2009](#_ENREF_20)). In this chapter, we analyze the theoretical framework of imbalanced classification, the main algorithmic approaches proposed in the literature and some of the most prominent applications in business. These business applications include customer relationship management (CRM) ([Kim, Chae, & Olson, 2012](#_ENREF_28)), fraud detection ([W. Wei, Li, Cao, Ou, & Chen, 2012](#_ENREF_52)), and risk management ([Groth & Muntermann, 2011](#_ENREF_17)).

**BACKGROUND**

Advances in science and technology accelerate the accessibility of raw data and create new opportunities for knowledge discovery. Imbalancedproblems can be found in a wide variety of applications, including security surveillance ([Wu, Wu, Jiao, Wang, & Chang, 2003](#_ENREF_54)), medical diagnosis ([Mena & JESUS, 2009](#_ENREF_35); [You, Zhao, Li, & Hu, 2011](#_ENREF_58)), bioinformatics ([Al-Shahib, Breitling, & Gilbert, 2005](#_ENREF_2)), geomatics ([Kubat, Holte, & Matwin, 1998](#_ENREF_29)), telecommunications ([Tang, Krasser, Judge, & Zhang, 2006](#_ENREF_47)), risk management ([Ezawa, Singh, & Norton, 1996](#_ENREF_16)), manufacturing ([Adam et al., 2011](#_ENREF_1)), quality estimation ([Lee, Song, Song, & Yoon, 2005](#_ENREF_30)), and power management ([Hu, Zhu, & Ren, 2008](#_ENREF_21)). Imbalanced classification has been studied in a number of studies ([N. V. Chawla, 2010](#_ENREF_10); [Guo, Yin, Dong, Yang, & Zhou, 2008](#_ENREF_18); [He & Garcia, 2009](#_ENREF_20); [Su, Mao, Zeng, Li, & Wang, 2009](#_ENREF_44); [Sun et al., 2009](#_ENREF_45)). Previous works on the classification of imbalanced data ([N. V. Chawla, 2010](#_ENREF_10); [Kubat et al., 1998](#_ENREF_29); [Ngai, Hu, Wong, Chen, & Sun, 2011](#_ENREF_36); [Su et al., 2009](#_ENREF_44); [Sun et al., 2009](#_ENREF_45)) address that many standard classification algorithms achieve poor performance. Therefore, despite the existing amounts of literature there is room for improvement and future contribution.

**MAIN FOCUS**

In this part, we present (1) the appropriate performance measures for imbalanced data; (2) imbalanced classification techniques and (3) the most popular business analytics applications.

**Performance Measures**

Classification performance measures can be obtained, directly or indirectly, from the confusion matrix. For a classification problem with classes, the confusion matrix is a square matrix , with each of its entries denoting the percentage of the samples that belong to the class and classified to the class . For the special case of binary classification (positive and negative), the confusion matrix is as follows:

Table 1: Confusion Matrix for binary classification problem

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Predicted | |
| Prevalent | Rare |
| Actual | Prevalent | TP  (True Positive) | FN  (False Negative) |
| Rare | FP  (False Positive) | TN  (True Negative) |

Clearly, the confusion matrix of an ideal classifier is diagonal. In this matrix, diagonal elements represent accurately classified examples and the off-diagonal elements the misclassified data for each class. A typical performance measure for classification is the so-called accuracy, which is calculated as the correctly classified samples over the total number of training samples. Since the majority class dominates the behavior of this metric, it might not be an appropriate performance indicator for imbalanced classification problems. More specifically a naive decision rule can yield high classification accuracy with no real practical value. For this, performance measures such as sensitivity and specificity are often employed:

Sensitivity value is driven by the correct classification of the majority class whereas specificity depends on the minority class. The plot of sensitivity versus specificity is called *Receiver Operator Characteristic* (ROC) curve and it provides a good visual representation of the classifier (Figure 1). A combined measure frequently used for imbalanceddata is the geometric mean of sensitivity and specificity (often abbreviated G-mean) defined by

There are other metrics used in the literature, including precision and recall or hit rate which is the ratio of true positive to the sum of true positive and false positive ([Duman, Ekinci, & Tanrıverdi, 2012](#_ENREF_14)) and lift which is highly related to accuracy, but it is well used in marketing practice ([Ling & Li, 1998](#_ENREF_34)). For a comprehensive review of classification performance measures we refer the reader to ([Sokolova & Lapalme, 2009](#_ENREF_43)).



Figure 1: ROC curve showing four classifiers

**Imbalanced Classification Techniques**

Several classification techniques have been proposed and applied in the literature for imbalanced classification problems. These techniques can be classified in two major categories:

* **Resampling** techniques are among the most popular preprocessing methods. Under this framework data points are added (oversampling) or removed (undersampling) to create a balanced problem. The Synthetic Minority Oversampling Technique (SMOTE) belongs to this category ([N. V. Chawla, Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002](#_ENREF_11)). However, resampling methods become inefficient for highly imbalanced problems with limited minority class examples and when data distribution are unknown ([Elazmeh, Japkowicz, & Matwin, 2006](#_ENREF_15)). In fact, oversampling often suffers from induced bias or overfitting, whereas through undersampling it is possible to lose valuable information by removing data.
* **Cost-sensitive learning algorithms** assign weights to data examples based on their importance. They are equivalent to resampling technique and combine both undersampling and oversampling. Many popular classification algorithms can be adapted under this framework. The SVM adaptation is termed weighted support vector machine(also termed Fuzzy SVM) which was originally proposed by ([Lin & Wang, 2002](#_ENREF_33)) and further applied and studied in subsequent works ([An & Liang, 2013](#_ENREF_3); [Ke, Liu, & Pan, 2013](#_ENREF_27); [Q. Zhang, D. Liu, Z. Fan, Y. Lee, & Z. Li, 2012](#_ENREF_59); [Qiongsheng Zhang, Dong Liu, Zhidong Fan, Ying Lee, & Zhuojun Li, 2012](#_ENREF_60)). Their advantage is that the cost coefficient is directly factored into the SVM problem providing an exact optimal solution. Assume that a dataset is represented by a set of data point J=where l and n are the number of samples and features, respectively, and each is a sample with n features and a class label . The costs for two classes (minority and majority) are represented with and The weighted SVM classifies the data points by identifying a separating hyperplane whose distance is maximum with respect to the data points of each class. The separation hyperplane defined by the parameters w and b can be obtained by solving the following convex optimization problem.

(1)

s.t. (2)

(3)

where is the kernel function where m , i.e. each training sample is mapped into a higher dimensional space by the function The slack variables are added to the objective function whose goal is to allow but penalize misclassified points.

Recursive cost sensitive classification have also been proposed that adapt the weights so that the training geometric mean is maximized ([Imam, Ting, & Kamruzzaman, 2006](#_ENREF_25)).

* + **Cost-Sensitive with Adaptive Boosting** is an ensemble learning classification scheme that combines multiple “weak” classifiers in a single “strong” classifier. Boosting algorithms is just another classification algorithm that can be adapted into the cost sensitive framework. The boosting algorithms such as Ada-boost performs learning iteratively and assigns weights to each example in a way the misclassified examples from previous learning are assigned with higher weights than the correctly classified examples ([Changrampadi, Yun, & Gu, 2012](#_ENREF_9); [Schapire, Freund, Bartlett, & Lee, 1998](#_ENREF_42)). Other ensemble learning method such as bagging ([Breiman, 2001](#_ENREF_5)) constructs an ensemble of multiple base classifiers by random uniformly sampling from the original training data set. Unlike boosting, all training examples are equally weighted in bagging methods.

**Applications in Business Analytics**

The class imbalance problem has a variety of applications in the business field. The following examples explain some specific problems in which imbalanced classification problems arise. We also discuss the related techniques to tackle the classification of imbalanced data in these problems.

* **Fraud detection.** The rate of fraud is growing extremely quickly along with the development of modern technology and communication, yielding the loss of millions of dollars each year. Coming up with a solution for this problem is tremendously important for numerous business associations. There are more reliable users than fraudulent ones in transaction databases. Organizations try to identify fraud by monitoring suspicious transaction patterns in their databases. There are different types of fraud, including credit card fraud, telecommunication or cellular fraud, online banking fraud, and insurance fraud. Each type has specific characteristics. For instance, the majority of online banking fraud problems deal with online banking transaction data sets with these characteristics and challenges: (1) highly imbalanced large data set; (2) real time detection, and (3) diverse customer behavior patterns. A range of techniques have been used in the literature for this problem, including a cost-sensitive neural network technique ([Wei Wei, Li, Cao, Ou, & Chen, 2013](#_ENREF_53)) and a random forests technique ([Bradter, Kunin, Altringham, Thom, & Benton, 2012](#_ENREF_4)). A cost-sensitive neural network is a modified neural network-based method by using weighting or scoring strategy. A random forest consisting of multiple decision trees is a modified version of the classical decision tree method. Also there are some techniques to address the credit card fraud detection problem, such as a combination of k reverse nearest neighbors (kRNN) method and sampling technique ([Padmaja, Dhulipalla, Bapi, & Krishna, 2007](#_ENREF_38)).There are few papers tackling insurance fraud detection. These techniques include, combined approach of minority oversampling with backpropagation (BP), naive bayesian (NB) and C4.5 algorithms ([Phua, Alahakoon, & Lee, 2004](#_ENREF_41)); hybrid undersampling approach along with kRNN and K-means algorithms ([Vasu & Ravi, 2011](#_ENREF_50)); cost-sensitive learning with decision tree, Bayesian network, and bagging of them ([Cao, Chen, Li, Wei, & Ou, 2013](#_ENREF_7)); the bagging and Metacost methods with various decision tree based learning algorithms as weak classifiers ([Perera, Neupane, Faisal, Aung, & Woon, 2013](#_ENREF_39)); and a cost-sensitive support vector machine and decision tree based algorithms ([Di Martino, Decia, Molinelli, & Fernández, 2013](#_ENREF_13)).
* **Customer relationship management (CRM).** Identifying potential contributors or customers is of great importance for a company’s sales, profits, and improvement. In recent years, academic researchers and customer data analyzers have focused on developing the related databases and data analysis techniques. In ([Olson, 2007](#_ENREF_37)), the applications of ([Yan Tu & Yang, 2012](#_ENREF_49))data mining in CRM is reviewed. However, there are some studies dealing with imbalanced CRM datasets ([Kim et al., 2012](#_ENREF_28); [Y. Tu, Yang, & Benslimane, 2011](#_ENREF_48)). In fact, most datasets in the real world are more likely to be imbalanced while a binary variable is used for prediction (i.e. 1 for purchase and 0 for no purchase). The proportion of 1 in the datasets is too few. The techniques commonly used to address these problems include the combined approach of SVM with undersampling ([Kim et al., 2012](#_ENREF_28)), a cost-based version of bayesian network classification ([Y. Tu et al., 2011](#_ENREF_48)), the combined ensemble learning with cost-sensitive learning ([Xiao, Xie, He, & Jiang, 2012](#_ENREF_56)), the bagging methods with decision tree algorithm along a feature reduction technique ([Yan Tu & Yang, 2012](#_ENREF_49)), and a weighted random forest ([Burez & Van den Poel, 2009](#_ENREF_6)). One of the interesting topics in CRM is churn prediction which has become one of the main challenges of many companies ([Chandar, Laha, & Krishna, 2006](#_ENREF_8); [B. Huang, Kechadi, & Buckley, 2012](#_ENREF_22)).
  + **Churn prediction.** Customer churn is the tendency of customers to terminate doing business with a company (e.g. bank, financial institution and so on) in a given period of time. Customer churn is a frequently rare event in service industries ([Gupta et al., 2006](#_ENREF_19)) but still important to detect. Different classification techniques have been applied for the imbalanced churn prediction problem, such as weighted random forest and logistic regression ([Burez & Van den Poel, 2009](#_ENREF_6)), a random forest technique together with the sampling techniques and cost-sensitive learning ([Xie, Li, Ngai, & Ying, 2009](#_ENREF_57)), hybrid undersampling approach along with KRNN and K-means algorithms ([Vasu & Ravi, 2011](#_ENREF_50)), a feature selection technique and RotBoost based ensemble classiﬁcation ([Idris, Khan, & Lee, 2013](#_ENREF_24)). A new hybrid method of clustering and classification data mining approaches ([Y. Huang & Kechadi, 2013](#_ENREF_23)) is proposed to deal with imbalanced telecomm churner data. They report the AUC and ROC to evaluate the prediction accuracy.

Table 2: Imbalanced classification problems in business applications

|  |  |  |  |
| --- | --- | --- | --- |
| Author(s)/  Year | Imbalanced classification technique | Business domain | Performance measures |
| ([Phua et al., 2004](#_ENREF_41)) | Hybrid oversampling with BP, NB, C4.5 | Fraud detection | Accuracy |
| ([Padmaja et al., 2007](#_ENREF_38)) | Hybrid KRNN and resampling | Fraud detection | True Positive rate and True Negative rate |
| ([Perols, 2011](#_ENREF_40)) | Logistic Regression, SVM, ANN, Bagging, C4.5, and Stacking | Fraud detection | Estimated Relative Costs of Misclassification (ERC) |
| ([Wei Wei et al., 2013](#_ENREF_53))  ([Di Martino et al., 2013](#_ENREF_13)) | Cost-sensitive ANN  Cost-sensitive SVM and Decision tree algorithms | Fraud detection  Fraud detection | Accuracy  F-measure, Recall rate |
| ([Y. Tu et al., 2011](#_ENREF_48)) | Cost-based version of Bayesian Network | CRM | AUC And Sensitivity |
| ([Kim et al., 2012](#_ENREF_28)) | SVM with Random Undersampling | CRM | Accuracy, Sensitivity, and Specificity |
| ([Burez & Van den Poel, 2009](#_ENREF_6)) | Random Undersampling, Gradient Boosting, and a Weighted Random Forest | Churn prediction | AUC and Lift |
| ([Xie et al., 2009](#_ENREF_57))  ([Idris et al., 2013](#_ENREF_24)) | Random Forest together with Resampling  RotBoost based ensemble | Churn prediction  Churn prediction | Lift  Sensitivity, Specificity, AUC |
| ([Duman et al., 2012](#_ENREF_14))  ([Xiao, He, & Wang, 2014](#_ENREF_55)) | Logistic regression, ANN, and the Chi-Squared automatic interaction detector (CHAID)  Neural network ensemble learning | Marketing  Marketing | Accuracy, AUC, and Precision (Hit rate)  Hit rate |
| ([Vasu & Ravi, 2011](#_ENREF_50)) | Hybrid undersampling with KRNN And K-Means algorithms | Insurance fraud detection and churn prediction | Sensitivity, Specificity, AUC, and Accuracy |
| ([Ezawa et al., 1996](#_ENREF_16))  ([Liao, Shih, Chen, & Hsu, 2014](#_ENREF_32)) | Bayesian network learning  Back propagation neural network ensemble learning | Risk management  Risk management | ROC  Accuracy, 1-specificity, 1-sensitivity |
|  |  |  |  |

* **Marketing.** For direct marketing tasks such as direct mail and catalog firms, accurate prediction of consumer feedback is essential for increasing profitability. Direct marketing classification has is an important classification problems in the business domain, and consists in classifying consumers into the buyers and non-buyers based on the predicted probabilities ([Cui, Wong, Zhang, & Li, 2008](#_ENREF_12); [Duman et al., 2012](#_ENREF_14); [Ling & Li, 1998](#_ENREF_34)). Methods applied in this problem include logistic regression, neural network and CHAID algorithms ([Duman et al., 2012](#_ENREF_14)). Customer targeting models have become surprisingly an important task in database marketing. Several researches developed classification data mining models for customer targeting ([Tan, Yeoh, Boo, & Liew, 2013](#_ENREF_46); [Xiao et al., 2014](#_ENREF_55)). A classifier ensemble model based on group method of data handling (GMDH) type neural network has been proposed by ([Xiao et al., 2014](#_ENREF_55)).
* **Risk management**. There are few studies addressing the classification problem of risk management datasets. The initial work done by ([Ezawa et al., 1996](#_ENREF_16)) implemented bayesian network model for predicting uncollectibles using imbalanced datasets in the telecommunication industry. A more recent work ([W. Wei et al., 2012](#_ENREF_52)) has implemented an online banking risk management system using a risk scoring method. In their system, a voting method is combined the scores from the contrast pattern mining, and a cost-sensitive neural network technique. A new ensemble learning technique of back propagation neural network is used for financial risk mining ([Liao et al., 2014](#_ENREF_32)) .

Table 2 summarizes some previous studies that addressed the imbalanced classification problems in business applications.

**FUTURE TRENDS**

The development of imbalanced classification techniques and measures in business applications has grown over the last few years. However, there are still some challenges in this area. In the future imbalanced classification techniques could be improved though unsupervised, semi supervised or game-theoretic approaches. More study is needed for classification of business information on the web with imbalanced data. Mobile telecommunication businesses should collect more information about the churn behavior of prepaid customers. Churn prediction will then become more consistent. Practitioners would benefit more if data sharing are publicly available and the description of corporate needs was better described. Developing efficient performance measures in business applications when data are highly imbalanced is a motivating area for future research. It would be more helpful if the proposed techniques can be applicable for other similar applications.

Based on the literature, an area that has received little attention on direct marketing is the class imbalance problem. There is still a need to further study the customer response predictive problem with imbalanced data.

**CONCLUSION**

The performance of standard classification techniques (such as SVM) is highly affected when the problem is imbalanced. Several approaches and performance measures have been established in the literature in order to overcome the class imbalance problems. The classification algorithms help quicken and automate the manual part of a screening/checking process. In this chapter, the main algorithmic approaches for imbalanced data are discussed, along with some of the most prominent applications in business. Many business datasets are imbalanced by nature. Therefore, it is extremely crucial to incorporate such imbalanced data into the classification task. The business applications mainly include customer relationship management classification tasks (e.g. prediction and classification of a customer group membership and churn prediction), fraud detection, and risk management.

**REFERENCES**

Adam, A., Chew, L.C., Shapiai, M.I., Jau, L.W., Ibrahim, Z., & Khalid, M. (2011). *A Hybrid Artificial Neural Network-Naive Bayes for solving imbalanced dataset problems in semiconductor manufacturing test process.* Paper presented at the Hybrid Intelligent Systems (HIS), 2011 11th International Conference on, (pp. 133-138).

Al-Shahib, A., Breitling, R., & Gilbert, D. (2005). Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics, 4*(3), 195-203.

An, Wenjuan, & Liang, Mangui. (2013). Fuzzy support vector machine based on within-class scatter for classification problems with outliers or noises. *Neurocomputing, 110*, 101-110.

Bradter, Ute, Kunin, William E, Altringham, John D, Thom, Tim J, & Benton, Tim G. (2012). Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. *Methods in Ecology and Evolution, 4*(2), 167–174.

Breiman, Leo. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications, 36*(3), 4626-4636.

Cao, Longbing, Chen, Jia Hang, Li, Jinjiu, Wei, Wei, & Ou, Yuming. (2013). Effective Detection Of Sophisticated Online Banking Fraud On Extremely Imbalanced Data. *World Wide Web*, 1-27.

Chandar, M., Laha, A., & Krishna, P. (2006). *Modeling churn behavior of bank customers using predictive data mining techniques.* Paper presented at the National conference on soft computing techniques for engineering applications (SCT-2006) (pp. 24-26).

Changrampadi, Mohamed H, Yun, Yixiao, & Gu, Irene YH. (2012). *Multi-class ada-boost classification of object poses through visual and infrared image information fusion.* Paper presented at the Pattern Recognition (ICPR), 2012 21st International Conference on.

Chawla, N.V. (2010). Data mining for imbalanced datasets: An overview. *Data Mining and Knowledge Discovery Handbook*, 875-886.

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

Cui, G., Wong, M.L., Zhang, G., & Li, L. (2008). Model selection for direct marketing: performance criteria and validation methods. *Marketing Intelligence & Planning, 26*(3), 275-292.

Di Martino, Matías, Decia, Federico, Molinelli, Juan, & Fernández, Alicia. (2013). A Novel Framework for Nontechnical Losses Detection in Electricity Companies *Pattern Recognition-Applications and Methods* (pp. 109-120): Springer.

Duman, E., Ekinci, Y., & Tanrıverdi, A. (2012). Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Systems with Applications, 39*(1), 48-53.

Elazmeh, W., Japkowicz, N., & Matwin, S. (2006). Evaluating misclassifications in imbalanced data. *Machine Learning: ECML 2006*, 126-137.

Ezawa, K.J., Singh, M., & Norton, S.W. (1996). *Learning goal oriented Bayesian networks for telecommunications risk management.* Paper presented at the Machine Learning-International Workshop Then Conference- (pp. 139-147). Morgan Kaufmann Publishers, Inc.

Groth, S.S., & Muntermann, J. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems, 50*(4), 680-691.

Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. ( 2008). *On the class imbalance problem.* Paper presented at the Natural Computation, 2008. ICNC'08. Fourth International Conference on (Vol. 4, pp. 192-201). IEEE.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., . . . Sriram, S. (2006). Modeling customer lifetime value. *Journal of Service Research, 9*(2), 139-155.

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

Hu, G.S., Zhu, F.F., & Ren, Z. (2008). Power quality disturbance identification using wavelet packet energy entropy and weighted support vector machines. *Expert Systems with Applications, 35*(1), 143-149.

Huang, Bingquan, Kechadi, Mohand Tahar, & Buckley, Brian. (2012). Customer churn prediction in telecommunications. *Expert Systems with Applications, 39*(1), 1414-1425.

Huang, Ying, & Kechadi, Tahar. (2013). An Effective Hybrid Learning System for Telecommunication Churn Prediction. *Expert Systems with Applications, 40*(14), 5635–5647.

Idris, Adnan, Khan, Asifullah, & Lee, Yeon Soo. (2013). Intelligent churn prediction in telecom: employing mRMR feature selection and RotBoost based ensemble classification. *Applied Intelligence*, 1-14.

Imam, T., Ting, K., & Kamruzzaman, J. (2006). z-svm: An svm for improved classification of imbalanced data. *AI 2006: Advances in Artificial Intelligence*, 264-273.

Japkowicz, N. (2000). *The class imbalance problem: Significance and strategies.* Paper presented at the Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI’2000) (Vol. 1, pp. 111-117).

Ke, Hong Xia, Liu, Guo Dong, & Pan, Guo Bing. (2013). Fuzzy Support Vector Machine for PolSAR Image Classification. *Advanced Materials Research, 639*, 1162-1167.

Kim, G., Chae, B.K., & Olson, D.L. (2012). A support vector machine (SVM) approach to imbalanced datasets of customer responses: comparison with other customer response models. *Service Business*, 1-16.

Kubat, M., Holte, R.C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine learning, 30*(2), 195-215.

Lee, D.E., Song, J.H., Song, S.O., & Yoon, E.S. (2005). Weighted support vector machine for quality estimation in the polymerization process. *Industrial & engineering chemistry research, 44*(7), 2101-2105.

Lemnaru, Camelia, & Potolea, Rodica. (2012). Imbalanced Classification Problems: Systematic Study, Issues and Best Practices. *Enterprise Information Systems*, 35-50.

Liao, Jui-Jung, Shih, Ching-Hui, Chen, Tai-Feng, & Hsu, Ming-Fu. (2014). An ensemble-based model for two-class imbalanced financial problem. *Economic Modelling, 37*, 175-183.

Lin, C.F., & Wang, S.D. (2002). Fuzzy support vector machines. *Neural Networks, IEEE Transactions on, 13*(2), 464-471.

Ling, C.X., & Li, C. (1998). *Data mining for direct marketing: Problems and solutions.* Paper presented at the Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (pp. 73-79).

Mena, Luis, & JESUS, A GONZALEZ. (2009). Symbolic one-class learning from imbalanced datasets: application in medical diagnosis. *International Journal on Artificial Intelligence Tools, 18*(02), 273-309.

Ngai, EWT, Hu, Yong, Wong, YH, Chen, Yijun, & Sun, Xin. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems, 50*(3), 559-569.

Olson, D.L. (2007). Data mining in business services. *Service Business, 1*(3), 181-193.

Padmaja, T.M., Dhulipalla, N., Bapi, R.S., & Krishna, P.R. (2007). *Unbalanced data classification using extreme outlier elimination and sampling techniques for fraud detection.* Paper presented at the Advanced Computing and Communications, 2007. ADCOM 2007. International Conference on (pp. 511-516). IEEE.

Perera, Kasun S, Neupane, Bijay, Faisal, Mustafa Amir, Aung, Zeyar, & Woon, Wei Lee. (2013). A Novel Ensemble Learning-Based Approach for Click Fraud Detection in Mobile Advertising *Mining Intelligence and Knowledge Exploration* (pp. 370-382): Springer.

Perols, J. (2011). Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms. *AUDITING: A Journal of Practice & Theory, 30*(2), 19-50.

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *ACM SIGKDD Explorations Newsletter, 6*(1), 50-59.

Schapire, Robert E, Freund, Yoav, Bartlett, Peter, & Lee, Wee Sun. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics, 26*(5), 1651-1686.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427-437.

Su, P., Mao, W., Zeng, D., Li, X., & Wang, F.Y. (2009). *Handling class imbalance problem in cultural modeling.* Paper presented at the Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on (pp. 251-256). IEEE.

Sun, Y., Wong, A.K.C., & Mohamed, S.K. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*(04), 687-719.

Tan, Ding‐Wen, Yeoh, William, Boo, Yee Ling, & Liew, Soung‐Yue. (2013). The impact of feature selection: A data-mining application in direct marketing. *Intelligent Systems in Accounting, Finance and Management, 20*(1), 23-38.

Tang, Y., Krasser, S., Judge, P., & Zhang, Y.Q. (2006). *Fast and Effective Spam Sender Detection with Granular SVM on Highly Imbalanced Mail Server Behavior Data.* Paper presented at the Collaborative Computing: Networking, Applications and Worksharing, 2006. CollaborateCom 2006. International Conference on (pp. 1-6). IEEE.

Tu, Y., Yang, Z., & Benslimane, Y. (2011). *Towards an optimal classification model against imbalanced data for Customer Relationship Management.* Paper presented at the Natural Computation (ICNC), 2011 Seventh International Conference on (Vol. 4, pp. 2401-2405). IEEE.

Tu, Yan, & Yang, Zijiang. (2012). An enhanced Customer Relationship Management classification framework with Partial Focus Feature Reduction. *Expert Systems with Applications, 40*(6), 2137-2146.

Vasu, M., & Ravi, V. (2011). A hybrid under-sampling approach for mining unbalanced datasets: applications to banking and insurance. *International Journal of Data Mining, Modelling and Management, 3*(1), 75-105.

Wang, L., Zhao, H., Dong, G., & Li, J. (2005). On the complexity of finding emerging patterns. *Theoretical Computer Science, 335*(1), 15-27.

Wei, W., Li, J., Cao, L., Ou, Y., & Chen, J. (2012). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 1-27.

Wei, Wei, Li, Jinjiu, Cao, Longbing, Ou, Yuming, & Chen, Jiahang. (2013). Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 1-27.

Wu, G., Wu, Y., Jiao, L., Wang, Y.F., & Chang, E.Y. (2003). *Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance.* Paper presented at the Proceedings of the eleventh ACM international conference on Multimedia (pp. 528-538). ACM.

Xiao, Jin, He, Changzheng, & Wang, Shouyang. (2014). *A Classifier Ensemble Model Based on GMDH-type Neural Network for Customer Targeting.* Paper presented at the Proceedings of the Seventh International Conference on Management Science and Engineering Management (pp. 259-269).

Xiao, Jin, Xie, Ling, He, Changzheng, & Jiang, Xiaoyi. (2012). Dynamic classifier ensemble model for customer classification with imbalanced class distribution. *Expert Systems with Applications, 39*(3), 3668-3675.

Xie, Y., Li, X., Ngai, EWT, & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications, 36*(3), 5445-5449.

You, M., Zhao, R.W., Li, G.Z., & Hu, X. (2011). MAPLSC: A novel multi-class classifier for medical diagnosis. *International journal of data mining and bioinformatics, 5*(4), 383-401.

Zhang, Q., Liu, D., Fan, Z., Lee, Y., & Li, Z. (2012). Feature and sample weighted support vector machine. *Knowledge Engineering and Management, 123*, 365-371.

Zhang, Qiongsheng, Liu, Dong, Fan, Zhidong, Lee, Ying, & Li, Zhuojun. (2012). Feature and sample weighted support vector machine. *Knowledge Engineering and Management*, 365-371.

**KEY TERMS AND THEIR DEFINITIONS**

**Contrast Pattern Mining:** An NP-hard pattern recognition method ([Wang, Zhao, Dong, & Li, 2005](#_ENREF_51)) to efficiently mine contrast patterns and separate fraudulent from genuine behavior.

**Imbalanced Data:** the data with different degrees of skewness between classes.

**Random Forest:** An ensemble learning technique by constructing decision tree ensembles**.**

**Semi Supervised Learning:** A[machine learning](http://en.wikipedia.org/wiki/Machine_learning) technique that uses both labeled and unlabeled [data](http://en.wikipedia.org/wiki/Data) for training.

**Supervised Learning:** A machine learning technique of predicting the value of a given function for any input based on labeled training data.

**Support Vector Machine (SVM):** A supervised machine learning method which analyzes data and distinguishes patterns for classification and regression analysis purposes based on convex optimization.

**Unsupervised Learning:** A machine learning technique of detecting unknown pattern in unlabeled data.

**Weighted Support Vector Machine (WSVM**): A modified type of SVM which assigns weights to different examples in the dataset.