A Benefit Optimization approach to the Evaluation of Classification Algorithms

Shellyann Sooklal and Patrick Hosein

Department of Computer Science, The University of the West Indies
St. Augustine, Trinidad
shellyann.sooklal@gmail.com, patrick.hosein@sta.uwi.edu

Abstract. We address the problem of binary classification when applied to non-communicable diseases. In such problems the data are typically skewed towards samples of healthy subjects. Because of this, traditional performance metrics (such as accuracy) are not suitable. Furthermore, classifiers are typically trained with the assumption that the benefit or cost associated with decision outcomes are the same. In the case of non-communicable diseases this is not necessarily the case since it is more important to err on the side of treatment of the disease rather on the side of over-diagnosis. In this paper we consider the use of benefits/costs for evaluation of classifiers and we also propose how the Logistic Regression cost function can be modified to account for these benefits and costs for better training to achieve the desired goal. We then illustrate the advantage of the approach for the case of identifying diabetes and breast cancer.

Keywords: Supervised binary classification · Health analytics · Performance measures · Benefit-based analytics

1 Introduction

Advances in biotechnology have lead to the production and access to a vast amount of data in health care. Medical records are now being digitized and digital machines and sensors contribute to the production of large amounts of data [14]. One of the most popular applications of this type of data is disease diagnosis. Some common machine learning classifiers which have been used for this task are Logistic Regression, Neural Networks, Support Vector Machines, Naive Bayes and also decision trees [8][18][21]. However, data on diseases are normally highly skewed towards the negative class; that is, the class associated with being healthy or disease free. Thus, this results in classifiers being biased towards the majority (negative) class and hence most, or in some cases all, of the minority (positive) class instances are incorrectly classified [8][18][16][21][15][22]. Also, the classifiers assume equal costs for different types of misclassification errors and thus reports high accuracy since the majority class would be correctly classified. However, in the real world, this is not desirable since the cost of misdiagnosing a positive instance is high. For example, misdiagnosing someone as healthy when they have

a disease can lead to earlier death or diminished quality of life if untreated. On the other hand, misdiagnosing a healthy person as having a disease can cause the person unnecessary stress and medical expenses. The consequences of the latter case are far fewer than the former therefore misclassification of these different types of errors should be associated with different costs. Similarly, the benefit of correctly classifying someone as positive for a disease is more valuable than correctly classifying a healthy person as healthy.

In this paper we provide a cost-sensitive approach to the disease classification task. However, we consider the case where the algorithm is trying to maximize benefit or reduce cost. We first associate the different types of classifications (true negative, false negative, false positive and true positive) with benefit values and then adjust the cost function of a Logistic Regression (LR) classifier to incorporate these values. We tested our approach using data on two common chronic non-communicable diseases, diabetes and breast cancer. We also formulated and calculated a benefit score for each test case and compared the results to that of the traditional case of maximizing accuracy. We also bench-marked our results against that of traditional LR and LR with SMOTE [4]. Hence, our contributions are a benefit-based LR algorithm as well as a benefit score for evaluating the performance of a classifier. In addition, we illustrate how one can derive such benefit scores.

The rest of the paper is organized as follows. Section 2 gives a brief description of previous approaches to classification with imbalanced data and traditional performance metrics. In section 3 we formulate a metric for benefit-based classification and then in section 4 we describe how it was incorporated into the LR classifier. Section 5 explains how the benefit-based metrics and classifier were applied to diabetes while section 6 considers the breast cancer case.

2 Related Work

Many researchers have already applied different techniques to deal with the issues posed by imbalanced data. Both [7] and [9] presented detailed summaries of the work performed by these researchers and described the main approaches to deal with imbalanced data and also common metrics to evaluate the performance of the classifier. Although, [9]'s summary is almost ten years old, the main points from the review are still relevant today and also matches up to that of [7]. Therefore, instead of looking at individual works, we highlight the main points from both [7] and [9] and focus on one previous work that is similar to ours.

Two popular approaches presented by [7] and [9] to handle imbalanced data are sampling methods and cost-sensitive methods. Sampling methods involve either removing samples in the majority class (under-sampling) or adding copies of instances of the minority class (oversampling) in order to reduce the imbalance of the data. Under-sampling poses the issue of losing important aspects of the majority class whilst oversampling introduces the issue of over-fitting due to replicated data. Some algorithms, such as EasyEnsemble and SMOTE, try to address these issues but they require multiple classifiers and additional processing

of the data which adds to the complexity and computation time of the solution. Also, there is still generalization and loss of information in these techniques.

On the other hand, cost-sensitive methods addresses the imbalanced data problem by assigning misclassification costs to incorrect classification. Costs can be applied to the training stage of the classification process where the objective is to minimize the overall cost. [7] found more papers implementing sampling methods than cost-sensitive methods; however, they explained that a reason for this could be due to sampling techniques being easier to implement since it does not required knowledge of the classification algorithms. On the other hand, [9] pointed out that studies have shown cost-sensitive learning to be more effective than sampling techniques and can therefore be used as a viable alternative. Thus, it was the chosen method for our purposes. A cost matrix is typically used to represent the penalties associated with misclassifying a sample. An example of the cost matrix for a binary classification is shown in Table 1.

Table 1. Cost-Matrix for Cost-Sensitive Classification

	Predicted No	**Predicted Yes**
Actual No	True Negatives, TN (Cost = 0)	False Positives, FP (Cost > 0)
Actual Yes	False Negatives, FN (Cost > 0)	True Positives, TP (Cost = 0)

A cost of zero is normally assigned to correct classification whereas non-zero values are assigned to misclassification. However, our proposed method goes a step further and assigns non-zero values for all four categories, unless it is required for the specific context, and hence, aims at maximizing benefit while reducing cost. Many of the proposed methods use complex classifiers or ensembles of classifiers in order to achieve reasonable results. However, our approach makes a simple change to a LR classifier in order to include benefit values into the learning process. Thus proving that satisfactory results can be achieved without adding complexity and computational costs.

Bahnsen et. al. [3] implemented a cost-sensitive LR algorithm for credit scoring. Their solution applied individual cost matrices to each example in the dataset. Their original formulation to include cost into the LR classifier is similar to ours. However, we go a step further in our formulation in order to account for both benefits and cost as well as represent these values as a single metric which influences the LR classifier. Also, manually specifying individual cost matrices for each sample in the dataset is very time consuming and impractical as the size of the dataset grows. This step also becomes impossible if the data was collected independently of the classification step. Our approach uses only one cost matrix that is generalized for all samples in the dataset and we show that this cost matrix can be easily formulated using readily available cost and benefit values.

Some of the most widely used performance metrics for classification are accuracy, precision, sensitivity, specificity, F1-score, receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) [7][6][3][20]. Accuracy, pre-

cision and F1-score are all dependent on data distribution and hence are not suitable for imbalanced data [9]. Recall and specificity are not data sensitive but they do not provide any indication of how many of the minority and majority samples were correctly classified. [9] explained that the ROC curve presents an optimistic view of the classifiers performance and do not provide any confidence intervals or statistical significance on the performance. [7] also stated that AUC/ROC can be incoherent.

The main issue with these traditional metrics is that they consider equal benefit for correct classification of each class and similarly, equal cost for misclassification. Therefore, when the data is skewed towards the majority class, these metrics illustrate high percentages for bias classifiers which in turn gives the impression that the classifier is performing well. Instead of using these standard metrics, [3] created a formula for calculating a savings score for credit scoring. Our work includes a benefit performance metric which is more suited for medical diagnosis and is also more generalized for other classification problems. Our benefit metric gives a better representation, than traditional metrics, of how the classifier is performing with both the majority and minority classes.

3 Benefit-Based Performance Metric

We focus on binary classification in which N labeled samples are provided. Each outcome belongs to either class 0 or class 1. If we denote the feature vector of a given sample by $\boldsymbol{x}$ then a classifier will produce a continuous score $s(\boldsymbol{x})$ which is used to determine the class in which it belongs. We assume that the instances for class 0 produce scores that are typically less than those of class 1. One must then determine some threshold t such that if $s(\boldsymbol{x}) \leq t$ the instance is classified as 0 while if $s(\boldsymbol{x}) > t$ then the instance is classified as a 1. For a given classifier, we denote the probability density function of the scores for class 0 instances by $f_0(s)$ and for class 1 scores by $f_1(s)$. We denote the corresponding cumulative distribution functions by $F_0(s)$ and $F_1(s)$ respectively.

Next we associate the corresponding costs and benefits. Let b_{ij} denote the benefit of classifying an instance of class i as class j. We assume that if $i = j$ (in which case the classification was correct) then the benefit is positive (i.e., $b > 0$) while if $i \neq j$ then the benefit is non-positive (or is rather a cost) (i.e., $b \leq 0$). The prior probability of class $j \in \{0, 1\}$ is denoted by π_j.

For a given threshold t, the expected number of instances from class 0 that are correctly classified can be written as $\pi_0 F_0(t) N$ which is the product of the probability that the instance is in class 0 times the probability that the instance is correctly classified (i.e., the score is less than the threshold t) times the total number of instances N. For a given threshold the expected benefit can then be written as

$$\begin{aligned} B(t) = \pi_0 F_0(t) b_{00} + \pi_0 (1 - F_0(t)) b_{01} + \\ \pi_1 F_1(t) b_{10} + \pi_1 (1 - F_1(t)) b_{11} \end{aligned} \tag{1}$$

Note that since only b_{00} and b_{11} are positive then the expected benefit is maximized when $F_0(t) = 1$ and $F_1(t) = 0$ which occurs when the two distributions

do not overlap. Therefore the benefit is upper bounded by $\pi_0 b_{00} + \pi_1 b_{11}$. If a classifier γ obtains an expected benefit of B_γ we define the following performance metric

$$\mu_\gamma \equiv \frac{B_\gamma}{\pi_0 b_{00} + \pi_1 b_{11}} \tag{2}$$

Note that if the classifier's performance is close to the optimal then $\mu_\gamma \approx 1$.

In general, for a given classifier and benefits we would like to maximize the benefit $B(t)$. A necessary condition for optimality can be obtained by taking the derivative of $B(t)$ with respect to t and setting the result to zero. If we do this then we obtain:

$$f_0(t^*)\pi_0(b_{00} - b_{01}) = f_1(t^*)\pi_1(b_{11} - b_{10}) \tag{3}$$

4 Benefit Objective with Logistic Regression

In the previous section we considered how, given the classifier, one could optimize the expected benefit by varying the threshold used by the classifier. However, the classifier itself was not trained with the objective of optimizing the benefit function. In this section we consider how the cost function used for LR can be modified to account for benefits and costs.

In LR, the posterior probability of the positive class is estimated as the logistic sigmoid of a linear function of the feature vector. For a given feature vector $\boldsymbol{x}_i$ this probability is given by

$$p_i = P(y = 1|\boldsymbol{x}_i) = h_\theta(\boldsymbol{x}_i) = g(\boldsymbol{\theta}^T \boldsymbol{x}_i) \tag{4}$$

where $h_\theta(\boldsymbol{x}_i)$ refers to the classification for input $\boldsymbol{x}_i$ given the parameter vector $\boldsymbol{\theta}$. The function $g(\cdot)$ is the logistic sigmoid function given by

$$g(z) = \frac{1}{1 + e^{-z}}. \tag{5}$$

The parameters $\boldsymbol{\theta}$ are determined by minimizing the LR cost function

$$J(\boldsymbol{\theta}) \equiv \frac{1}{N} \sum_{i=1}^{N} J_i(\boldsymbol{\theta}) \tag{6}$$

where

$$J_i(\boldsymbol{\theta}) = -y_i \log(h_\theta(\boldsymbol{x}_i)) - (1 - y_i) \log(1 - h_\theta(\boldsymbol{x}_i)) \tag{7}$$

However this cost function assumes the same cost is associated with different errors (false positives and false negatives). Consider, instead, the following cost function based on maximizing benefits b_{ij}.

$$J^B(\boldsymbol{\theta}) \equiv \frac{1}{N} \sum_{i=1}^{N} J_i^B(\boldsymbol{\theta}) \tag{8}$$

where

$$\begin{aligned} J_i^B(\boldsymbol{\theta}) = \; & y_i[h_\theta(\boldsymbol{x}_i)b_{11} + (1 - h_\theta(\boldsymbol{x}_i))b_{10}] + \\ & (1 - y_i)[h_\theta(\boldsymbol{x}_i)b_{01} + (1 - h_\theta(\boldsymbol{x}_i))b_{00}]. \end{aligned} \tag{9}$$

We can rewrite this as

$$\begin{aligned} J_i^B(\boldsymbol{\theta}) = \; & y_i h_\theta(\boldsymbol{x}_i)(b_{11} - b_{10}) + b_{10} \\ & (1 - y_i)(1 - h_\theta(\boldsymbol{x}_i))(b_{00} - b_{01}) + b_{01} \end{aligned} \tag{10}$$

Since we will be maximizing this function with respect to $\boldsymbol{\theta}$ and b_{10} and b_{01} are constants then they can be removed from the function without affecting the optimal $\boldsymbol{\theta}$. Furthermore we can change this into a minimization problem by multiplying by -1. Finally we can divide the resulting function by $(b_{11} - b_{10})$ and not change the optimal solution. If we do this we obtain the new function as

$$J_i^B = -y_i h_\theta(\boldsymbol{x}_i) - (1 - y_i)(1 - h_\theta(\boldsymbol{x}_i))\eta \tag{11}$$

where we have defined

$$\eta \equiv \frac{b_{00} - b_{01}}{b_{11} - b_{10}}. \tag{12}$$

Note that this has a similar form to the LR cost function but here the error for class 0 instances are scaled by the factor η. We therefore propose using the Logistic Cost function but include this scaling so that we take into account benefits. Hence we use the following:

$$J_i(\boldsymbol{\theta}) = -y_i \log(h_\theta(\boldsymbol{x}_i)) - \eta(1 - y_i) \log(1 - h_\theta(\boldsymbol{x}_i)) \tag{13}$$

when training the LR algorithm.

5 Application to Diabetes Classification

In this section we consider the problem of diabetes classification based on a number of patient features. We first determined appropriate benefit values and used LR with the computed value for η. We compared the results with traditional LR and also LR with SMOTE since SMOTE is a popular method for handling unbalanced data. We also included a wide range of other η values to illustrate the robustness of the approach. We then compared the results based on the benefits based optimization with those obtained with accuracy based optimization.

5.1 Dataset Description

The Pima Indian Diabetes dataset [13][19] was chosen for our experiments. The data contains 9 attributes based on measurements that are normally used to diagnose a patient (for example, BMI, age). It also contains 768 instances (500 for healthy persons, 268 for diabetic persons). 75% of the dataset was randomly selected for the training set and the remaining 25% was used for testing.

5.2 Cost-Based Benefit Model

We determine the benefits b_{ij} which are then used to compute η. We do this by determining benefits with respect to a baseline. The baseline case is the one in which nothing is done for the patient (i.e., as if the person was never a patient). The benefit b_{00} corresponds to the case when the patient has no diabetes and this was predicted correctly. In this case nothing will be done for the patient and so we set $b_{00} = 0$ since no benefit (except for the fact that the patient now knows they do not have diabetes) is achieved. The cost b_{10} corresponds to the case in which the patient has diabetes but it was not predicted. Now note that the baseline is the case where nothing would have been done for the patient anyway so the benefit (when compared to doing nothing) is still zero and so $b_{10} = 0$. Next consider the benefit b_{11} which is the case that the patient has diabetes and this was predicted and so treatment is provided to avoid health complications due to diabetes (which would have occurred in the baseline case). From [12], the annual cost of treatment for a patient with diabetes complications is approximately \$10,000 more than someone who did not have diabetes (or was treated early). Therefore the benefit of this correct detection of diabetes is a savings of \$10,000 so we set $b_{11} = 10$. Finally in the case b_{01} the patient was incorrectly diagnosed as having diabetes when they in fact did not have it. In this case drugs would be prescribed for the patient when they did not need it. Using information from [10], the annual cost of drugs for diabetes treatment for a new patient would be approximately \$1000 and so we use this as the associated cost (since the baseline would not have incurred this cost). Therefore we set $b_{01} = -1$. If we substitute these numbers we obtain $\eta = 0.1$.

What this means is that, when training, errors for the common instance (the patient does not have diabetes) are treated more lightly than errors for those with diabetes. According to [2] the prevalence of diabetes (in the United States) is typically on the order of 10% and hence we can approximate $\pi_0 = 0.9$ and $\pi_1 = 0.1$. We can now use these values to determine our proposed metric 2 as: $\mu = \frac{B}{1}$ where B is the expected benefit obtained and the upper bound is 1.

We can note the following. If a classifier always chooses 0 (no diabetes) then the accuracy will be 90% which is quite high. However the benefit cost for this classifier is 0 while the maximum possible benefit is 1. If a classifier always chooses 1 (has diabetes) then the accuracy now becomes 10% but the expected benefit becomes 0.1. This clearly shows the advantage of considering expected benefit rather than other non-cost dependent approaches.

5.3 Life Expectancy Based Benefit Model

In addition to medical expenses, we can determine benefits based on life expectancy. In the baseline case where nothing is done for a patient without diabetes, their life expectancy does not change, therefore we can set $b_{00} = 0$. However, if we do nothing for a person who has diabetes, that is, they are not diagnosed as being diabetic, then according to [11] and [5] their life expectancy can be reduced by 10 years. Therefore, we set $b_{10} = -10$. [5] also states that

diabetics are normally diagnosed years after developing the condition. Based on this fact, we set $b_{11} = 5$ to represent extending a person's life by at least half the amount of not detecting it at all. For the case where the person does not have diabetes but we predicted them as having it, diabetes drugs can be administered. [1] explained that overdose of the insulin taken by type 1 diabetics and also insulin-requiring type 2 diabetics can result in coma or even death. Therefore, we set $b_{01} = -1$ to represent a possibility of negative side effects (including affected lifespan) through the use diabetic medication. Hence using this approach we obtain $\eta = 0.07$ which is close to the value obtained with the cost based mode.

5.4 Numerical Results

Table 2 shows the results obtained for accuracy and benefit score for our models compared to regular LR and LR with SMOTE. For both models, our adjusted algorithm produced results which achieved the highest benefits; however, accuracy was sacrificed in order to get better performance with the minority class.

Table 2. Accuracy and Benefits achieved for LR, LR with SMOTE, Benefit-Based LR using Cost-Based Model and Benefit-Based LR using Life-Expectancy Model

Algorithm	**Accuracy**	**Cost-Based Benefit**	**Life-Based Benefit**
LR	0.81	2.02	-0.27
LR with SMOTE	0.74	2.52	0.54
BB-LR (Cost-Based Model)	0.34	2.67	NA
BB-LR (Life-Based Model)	0.34	NA	1.01

In Figure 1 we plot the performance metric 2 as a function of η. Note that at $\eta = 1$ we obtain the value that we would have obtained using the traditional approach of maximizing accuracy. Note that optimizing with respect to benefit always provides a better expected benefit than optimizing with respect to accuracy.

In Figure 2 we plot the accuracy as a function of η. Here we find that to achieve improved benefit we must sacrifice accuracy. In other words we should err more on the side of large benefits. Here we find that the accuracy is either the same as or worse than that obtained with the traditional approach of maximizing accuracy.

5.5 Discussion

As shown in Table 2, we can achieve higher benefits by applying η to the LR classifier to account for benefits and misclassification costs. This supports the fact that a simple adjustment can be made to the LR classifier to reduce the

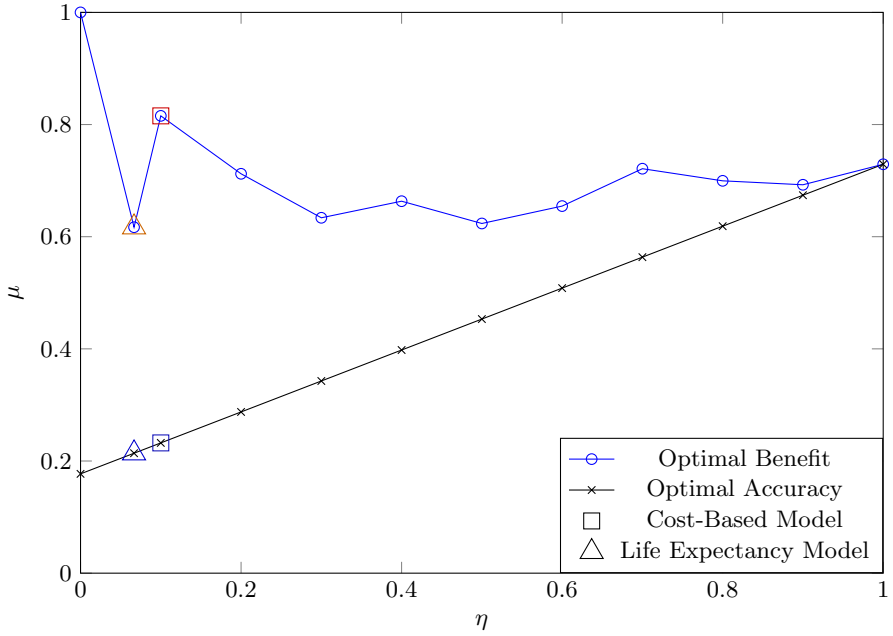

Fig. 1. Benefit Scores for Pima Indians Diabetes Dataset

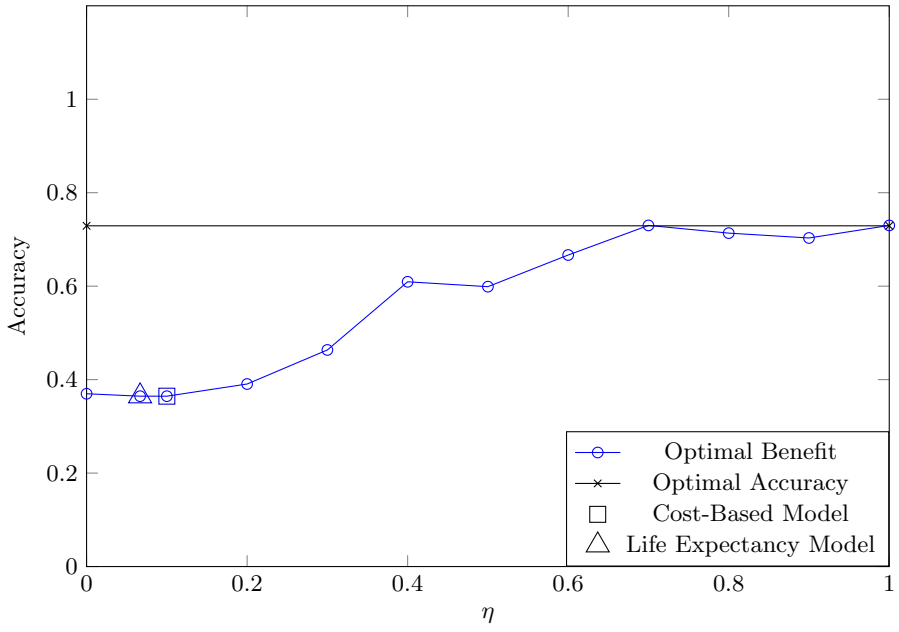

Fig. 2. Accuracy Scores for Pima Indians Diabetes Dataset

misclassification of the higher cost class. Thus, reducing complexity and computational costs which are introduced by other methods which handle imbalanced data such as SMOTE.

As explained in Section 5.4 above, the benefit-based LR classifier was tested against multiple benefit values and hence multiple values of η. When $\eta = 1.0$, this is identical to setting the classifier to perform accuracy optimization. Hence, as illustrated in Figures 1 and 2, the benefit and accuracy scores were the same for this case. Furthermore, for the derived value of $\eta = 0.1$, we obtained $B = 3.02$, $\mu = 0.82$, benefit-wrt-accuracy = 0.23 and accuracy = 0.36. From these values we see that optimizing with respect to benefit did indeed provide a better performance score than with the accuracy optimization approach. As illustrated in Figure 1, there was a difference of 0.59 and hence an increase of 256.52% in the performance value with respect to benefit than with the accuracy approach. Thus, the benefit approach can save a person on average $2,165.00 more per year than the accuracy approach which only saves a person approximately $850 per year. The accuracy score is very low (36%) and does not truly represent the performance of the classifier with respect to its classification on positive instances. For this scenario, the value of η was very small. This in turn represented a much higher benefit score for the positive class as opposed to the negative class. Therefore, the classifier would have highly skewed the classification to favor the positive class in order to maximize benefit. The performance value was high (82%) since the classification results obtained comprised of the majority of the positive class instances being classified as positive (true positives) but it also comprises of a high number of false positives. These results illustrates how sensitive the η value is in affecting the performance of the classifier.

Similarly, for the second derived value of $\eta = 0.07$ we obtained $B = 1.14$, $\mu = 0.62$, benefit-wrt-accuracy = 0.21 and accuracy = 0.36. The performance metric is slightly lower than than that of the previous case. This can be due to the benefit matrix associated with this value of η. However, when compared to the benefit based on accuracy, the benefit increases by 0.41 or 195.24%, as shown in Figure 1. This is an improvement in life expectancy by approximately 9 months per person over the accuracy approach which gives a life expectancy of approximately 4.6 months per person. The traditional accuracy value is very low for this scenario (36%), therefore, it does not provide a true reflection of the savings generated by the classifier. From these observations, the benefit approach is more robust than the accuracy approach since it gives a better classification based on benefits and it also gives a more accurate depiction of the performance of the classifier with respect to benefit.

6 Application to Breast Cancer Classification

Similar tests and derivation of η values were performed on a breast cancer dataset (the Breast Cancer Wisconsin (Diagnostic) dataset [17]), as with the diabetes dataset. Also, similar trends were observed in the results. However, due to space limitations we only provide the benefit values used for the experiments.

Table 3. Benefit Values for Cost-Based and Life-Expectancy Models for Breast Cancer

Model	b_{00}	b_{01}	b_{10}	b_{11}	η
Cost-Based	0	-30.5	0	74	0.41
Life-Expectancy	0	-1	-5	5	0.1

7 Conclusions

Medical datasets are typically skewed towards the negative class (healthy persons). This imbalance in the dataset influences classifiers to be biased towards the majority class. However, misclassification of a disease-affected person is more costly than the misclassification of a healthy person. In addition, the benefits associated with the correct classification of a disease-affected person is higher than that of a healthy person. In this paper, we presented a robust benefit-based LR approach which was sensitive to varying benefits and costs associated with different datasets/diseases. We also presented a benefit score which was used to evaluate the performance of the classifier. This score was successful in illustrating how the classifier performed with regards to overall gain in benefit. We are currently working on extending this work by incorporating benefit-based approaches to other classifiers.

References

1. A. Gundgurthi, S. Kharb, M.D.R.P., Garg, M.: Insulin poisoning with suicidal intent. Indian Journal of Endocrinology and Metabolism **16(Suppl1)**, S120 – S122 (2012). https://doi.org/10.4103/2230-8210.94254, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3354941/
2. Association, American Diabetes.: Statistics about diabetes (Jul 2017), http://www.diabetes.org/diabetes-basics/statistics/
3. Bahnsen, A.C., Aouada, D., Ottersten, B.: Example-dependent cost-sensitive logistic regression for credit scoring. In: Proceedings of the 2014 13th International Conference on Machine Learning and Applications. pp. 263–269. ICMLA '14, IEEE Computer Society, Washington, DC, USA (2014). https://doi.org/10.1109/ICMLA.2014.48, http://dx.doi.org/10.1109/ICMLA.2014.48
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: Synthetic minority over-sampling technique. J. Artif. Int. Res. **16**(1), 321–357 (Jun 2002), http://dl.acm.org/citation.cfm?id=1622407.1622416
5. Diabetes.co.uk: Diabetes life expectancy, https://www.diabetes.co.uk/diabetes-life-expectancy.html
6. Garrido, F., Verbeke, W., Bravo, C.: A robust profit measure for binary classification model evaluation. Expert Systems with Applications **92**, 154 – 160 (2018). https://doi.org/https://doi.org/10.1016/j.eswa.2017.09.045, http://www.sciencedirect.com/science/article/pii/S0957417417306498
7. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data. Expert Syst. Appl.

73(C), 220–239 (May 2017). https://doi.org/10.1016/j.eswa.2016.12.035, https://doi.org/10.1016/j.eswa.2016.12.035
8. He, F., Yang, H., Miao, Y., Louis, R.: A cost sensitive and class-imbalance classification method based on neural network for disease diagnosis. In: 2016 8th International Conference on Information Technology in Medicine and Education (ITME). pp. 7–10 (Dec 2016). https://doi.org/10.1109/ITME.2016.0012
9. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (Sept 2009). https://doi.org/10.1109/TKDE.2008.239
10. Helper, Health Cost.: How much does diabetes medication cost? (Oct 2013), http://health.costhelper.com/diabetes-medication.html
11. Huizen, J.: Type 2 diabetes and life expectancy (May 2017), https://www.medicalnewstoday.com/articles/317477.php
12. Institute, Health Care Cost.: Issue brief 10: Per capita health care spending on diabetes: 2009-2013 (May 2015), http://www.healthcostinstitute.org/files/HCCI Diabetes Issue Brief 205-7-15.pdf
13. Kaggle.com: Pima indians diabetes database, https://www.kaggle.com/uciml/pima-indians-diabetes-database/data
14. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I.: Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal **15**, 104 – 116 (2017). https://doi.org/https://doi.org/10.1016/j.csbj.2016.12.005, http://www.sciencedirect.com/science/article/pii/S2001037016300733
15. Krawczyk, B., Schaefer, G., Woniak, M.: A cost-sensitive ensemble classifier for breast cancer classification. In: 2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics (SACI). pp. 427–430 (May 2013). https://doi.org/10.1109/SACI.2013.6609012
16. Li, L., Chen, M., Wang, H., Li, H.: Cosfuc: A cost sensitive fuzzy clustering approach for medical prediction. In: 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery. vol. 2, pp. 127–131 (Oct 2008). https://doi.org/10.1109/FSKD.2008.378
17. Repository, U.M.L.: Breast cancer wisconsin (diagnostic) data set (Nov 1995), https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
18. Santos-Rodrguez, R., Garca-Garca, D., Cid-Sueiro, J.: Cost-sensitive classification based on bregman divergences for medical diagnosis. In: 2009 International Conference on Machine Learning and Applications. pp. 551–556 (Dec 2009). https://doi.org/10.1109/ICMLA.2009.82
19. Smith, J.W., Everhart, J., Dickson, W., Knowler, W., Johannes, R.: Using the adap learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the Annual Symposium on Computer Application in Medical Care. pp. 261–265 (November 1988)
20. Verbraken, T., Verbeke, W., Baesens, B.: A novel profit maximizing metric for measuring classification performance of customer churn prediction models. IEEE Transactions on Knowledge and Data Engineering **25**, 961–973 (2013)
21. Zhang, D., Shen, D.: Multicost: Multi-stage cost-sensitive classification of alzheimer's disease. In: Suzuki, K., Wang, F., Shen, D., Yan, P. (eds.) Machine Learning in Medical Imaging. pp. 344–351. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
22. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering **18**(1), 63–77 (Jan 2006). https://doi.org/10.1109/TKDE.2006.17