2007

Using Asymmetric Classification Cost Matrices in Predicting Diabetes

Bishwadip Ghosh
University of Colorado, Denver, bghosh@yahoo.com

Joseph Hasley
Health sciences Center, Denver, jhasley@hotmail.com

Using Asymmetric Classification Cost Matrices in Predicting Diabetes

Bishwadip Ghosh and Joseph Hasley

The Business School
University of Colorado at Denver and Health sciences Center
Denver, Colorado, United States of America
bghosh@yahoo.com, jhasley@hotmail.com

Abstract. Often there is a need to introduce classification costs into the classifier for predicting disease. This is determined by the type of disease, its associated classification cost matrix and/or the target population on which the classifier will be used. Diabetes has higher costs associated with false negatives than true positives, as the disease can progress very rapidly when left untreated. There are two ways to skew a classifier to work towards the given classification cost matrix: (1) by changing the classification probability value, P* based on the classification cost matrix or (2) by rebalancing the training set to introduce more negative cases. Using a diabetes data set, this paper compares the two methods. The results indicate comparable values of predictive accuracy and expected classification costs for either method. However, P* works better when the p-value is less than 0.2. Hence for diabetes classification matrices, the P* method is recommended.

Keywords: Classification, cost-sensitive learning, rebalancing.

1 Introduction

The Pima Native American diabetes dataset was collected by the National Institute of Diabetes and Digestive and Kidney Diseases and was donated for public use by Sigillito [1]. The data reports the diabetic status (diabetic or non-diabetic) of 768 women, along with data for 8 health-status variables. The population lives near Phoenix, Arizona, USA. The database has 768 cases (500 no disease, 268 disease). The data does not include classification costs. The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care.

Several studies have used the Pima Native American diabetes dataset to test the validity of various classifiers. [2] used the Pima diabetes data to test the performance

of a Generalized Partial Least Square classifier that they developed in comparison to six other commonly used classification algorithms. [3] developed a Bayesian logistic regression model and reported modest success when they tested it on the Pima data. Their results supported the use of the complete set of variables in the Pima data as well as a subset of the variables that produced similar error rates. [4] used the Pima data to compare their large margin nearest neighbor algorithm (LAMANNA) to yet another set of popular algorithms.

All of the above studies reported significant success differentiating diabetic and non-diabetic subjects in the Pima database. Indeed, when classification costs are symmetric (that is, when the costs of a false negative classification is roughly equivalent to the cost of a false-positive classification), accuracy, sensitivity, and specificity are appropriate measures of success.

2 Asymmetric Classification Costs for Diabetes

Classification costs are defined as the cost of correctly or incorrectly predicting a case as positive or negative. Given a specification of costs for correct or incorrect predictions, a case should be predicted based on the lowest expected cost [5]. Asymmetric classification cost matrices arise when the cost of true positive, false positive, true negative and false negative are different. These situations are common in certain domains, such as medical diagnostics. The resulting classification cost matrices are often asymmetric. In the instance of diabetes, for instance, a patient who is wrongly told that they are non-diabetic may go un-treated for years and incur irreversible cardio-vascular damage. Ultimately, the financial and quality of life costs of a false-negative diagnosis may prove exponentially greater than the costs of an erroneous false-positive. Hence in diabetes detection, the classifications costs for a false negative may be much higher than for a false positive, when all types of "costs" are considered [6]. [7] demonstrated and compared several cost-sensitive algorithms across various data-sets (including the Pima diabetes data-set.) The boosting procedures analyzed by Ting were based on the well known tree-learning algorithm C4.5 [8]. This study uses a regression-based method to examine the performance of Elkan's [5] rebalancing algorithm.

Moreover, the occurrence rate of diabetes varies greatly across different ethnicity and races. If a classifier for diabetes prediction is built from a sample of predominantly Anglo origin, to utilize that classifier on Asian subjects, classification costs need to be introduced to bias the classifier for the higher prevalence rates of diabetes seen in the Asian population.

2.1 Example Diabetes Costs Matrices

It is possible that for certain conditions, such as for high risk groups such as Asian populations or Native Indian population, it is better to have a false positive diagnosis than a true negative. Evidence suggests that some early treatment can improve physician performance, cost-effectiveness and patient outcomes [9]. This is because diabetes tends to be a silent disease for an extended period of time. Hence, even

before full blown diabetes is developed and diagnosed, a patient in the high risk domain may have severe consequences. Hence for that domain, the cost of a false positive may be lower than a true negative.

Reasonable diabetes classification cost matrices may involve the ratio of the cost of a false negative over the cost of a false positive set at 4 times, 9 times or 16 times, etc. The detection of diabetes also involves treatment costs and hence the ration of the cost of a true positive over the cost of a true negative can be 2, 3, 4 , etc. Table 1 shows some p* values generated from reasonable cost asymmetric cost matrices [10].

Table 1: p* values for Example Asymmetric Cost Matrices for Diabetes.

Cost of TN	Cost of FP	Cost of TP	Cost of FN	p* Value
1	2	4	8	0.2
1	3	9	27	0.1
1	4	16	64	0.058824
1	2	10	20	0.090909
1	2	10	30	0.047619
1	2	10	40	0.032258
1	2	10	50	0.02439
1	2	10	60	0.019608
1	2	20	60	0.02439
1	2	10	70	0.016393

Regardless of the actual scenarios, it may be noted that the p* value for typical diabetes classification cost matrices have very low p* values, between 0.016 to 0.20. Hence it is important to identify the optimal approach – whether P* or rebalance to introduce classification costs into a diabetes classifier.

3 Methods of Introducing Classification Costs

The two ways to bias the classifier to account for the asymmetric classification cost matrices are (Elkan, 2001):

1. Changing the classification probability value, P* based on the classification cost matrix. The classifier can be biased by using a different probability value, p*, which is calculated from the specified classification cost matrix using the formula (Elkan, 2001):

$$p^* = (C_{FP} - C_{TN}) / (C_{FP} - C_{TN} + C_{FN} - C_{TP}) . \quad (1)$$

   Where, C_{TP} is the cost of correctly classifying a case as positive and C_{TN} is the cost of correctly classifying a case as negative, C_{FP} is the cost of incorrectly classifying a case as positive and C_{FN} is the cost of incorrectly classifying a case as negative.
2. Rebalancing the training set to adjust the negative cases to bias the classifier to work towards the target p-value based on the asymmetric classification

cost matrix. The formula to make a target probability, P*, the number of negative cases in the training set needs to be multiplied by (Elkan, 2001):

$$\text{Number of negative cases} = P^* (1 - P_0) / (1 - P^*) (P_0) . \quad (2)$$

Where, P_0 is the probability threshold for the cost unaware classifier.

4 Research Goals

The goals of this research paper are as follows:

1. Evaluate the impact of introducing bias into the logistic regression classification algorithm by changing the p-value, henceforth, to be referred to as the P*-method and by rebalancing the sample. The specific measures to assess classification performance will be accuracy, specificity and sensitivity.
2. Understand whether one of the above methods will work better with the type of data being used – i.e., the diabetes data set.
3. Develop some understanding on the nature of diabetes data and the specific characteristics of the classification cost matrices that are prevalent in the domain of diabetes classification.

5 Research Hypotheses

When symmetric cost matrices are used an unbiased classifier is generated with a p-value of 0.5, without regard to the classification cost matrix. An unbiased classifier assumes the cost of a false positive is equal to the cost of a false negative and the cost of a true positive is equal to the cost of a true positive. This may not reflect the real world for scenarios such as classification of diabetes, where the classification costs are not equal. A biased classifier is developed by giving consideration to the classification cost matrix, hence the expected classification costs produced by the biased classifier (generated either by p* or rebalance methods) will be different from the unbiased classifier.

> *Hypothesis 1 - The mean of the expected classification costs for the unbiased classifier will be significantly different than the expected classification costs for the biased classifiers based on p* or rebalancing.*

Elkan (2001) contends that there is no significant difference in classification performance in terms of predictive accuracy and expected classification cost whether bias is introduced in the classifier by the P* or the rebalance methods given the same classification cost matrix.

Hypothesis 2 – The mean of the predictive accuracy for the biased classifiers using the p method and the rebalance methods will not be significantly different for different classification cost matrices.*

Hypothesis 3 - The mean of the expected classification costs for the biased classifier using p will not be significantly different from the mean of the expected classification costs for the biased classifier using rebalancing.*

As the p-value corresponding to the asymmetric classification cost matrix approaches the value of 0.2 or lower, the corresponding rebalancing ratio for negative cases becomes smaller and smaller. The result is that the number of negative cases in the training sample may become extremely small. Similarly, on the other extreme, as the p-value corresponding to the asymmetric classification cost matrix approaches the value of 0.8 or higher, the corresponding rebalancing ratio for negative cases becomes larger and larger. The result is that the number of positive cases in the training sample may become extremely small. In either of these two scenarios, the biased classifier will over learn the small number of cases and the end result will be poorer predictive accuracy for the biased classifier produced through rebalance over the one that produced through the p* method.

Proposition 1 - The performance of the biased classifier using p will be better than the biased classifier using rebalancing, when the p-value, corresponding to the asymmetric classification cost matrix is <= 0.2 or >= 0.8.*

6 Research Methodology

The methodology consists of comparing the performance of using the P* formula and the rebalancing formula from Elkan (2001) to introduce bias into the diabetes classifier to map different types of classification costs matrices. The cost matrices are developed to match the p* values listed in Table 1.

Experimental design will consist of the use of paired T-tests to compare the values of the performance measures across the two methods used to introduce classification cost bias – p* and rebalance. To establish statistical power for the experiment a total of 19 trials will be conducted evenly spread across the range of p* values (.05 to .95). The performance measures that will be compared include accuracy, specificity and sensitivity measures and expected classification costs.

6.1 Pima Diabetes Dataset

The data set consisted of 768 cases (500 with no disease and 268 with disease) with the following variables for each case:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

6.2 Experimental Steps

The following steps were performed. All logistic regressions were run using 20 fold re-sampling to improve algorithm accuracy and performance.

1. Cleaning the data to remove cases that have invalid values. This includes taking out the cases that have BMI (Body Mass Index) or glucose values of 0 and/or Blood Pressure of 0-10. There were 724 cases (475 with no disease and 249 with disease) in the final data set after cleanup.
2. The unbiased classifier was produced using a training/test split of 469 (301 with no disease and 168 with disease) in the training set and 255 (174 with no disease and 81 with disease) in the test set using Logistic Regression to classify the data without introducing any bias using a p value of 0.5. The accuracy, specificity and sensitivity measures for the classification were obtained from SPSS. The expected classification cost using a representative classification cost matrix was also obtained
3. The logistic regression was rerun with different P values from .05 to .95 at .05 intervals. The accuracy, specificity and sensitivity measures for the classification were obtained from SPSS. The expected classification cost for each run using a representative classification cost matrix was obtained.
4. The rebalance multiplier was calculated given a value of p* based on the classification cost matrix. The multiplier was used to adjust the negative training cases. SPSS was used to find a given number of random negative cases from the 301 negative cases in the training set.
5. Hypothesis testing was done using t-tests at alpha levels of .95 to check for significant differences in the means of the accuracy, sensitivity, specificity and expected classification costs from the rebalance method to the P* methods.

7 Results

The unbiased classifier for p=0.5 using the data set was generated using Logistic Regression and the Enter Method. The results with all the variables indicated that Blood Pressure, skinfold, times pregnant and insulin were not significant. The data

was cleaned and a subsequent logistic regression was run without skinfold, insulin[1] and blood pressure (see table 2).

The final regression coefficients and the classification results of the unbiased classifier are shown in Figure 1.

Table 2. Variable(s) entered on step 1: timesPreg, BMI, Pedigree, Age, PlsGlucose.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1(a)	timesPreg	.096	.047	4.192	1	.041	1.101
	BMI	.085	.018	23.334	1	.000	1.088
	Pedigree	.930	.355	6.851	1	.009	2.535
	Age	.011	.011	1.083	1	.298	1.011
	PlsGlucose	.031	.004	52.167	1	.000	1.031
	Constant	-8.370	.863	94.134	1	.000	.000

Classification Table(c)

	Observed		Predicted					
			Selected Cases(a)			Unselected Cases(b)		
			CLASS		Percentage Correct	CLASS		Percentage Correct
			0	1		0	1	
Step 1	CLASS	0	264	37	87.7	161	13	92.5
		1	78	90	53.6	32	49	60.5
	Overall Percentage				75.5			82.4

Figure 1. The unbiased Logistic Regression Model and classification results from SPSS

7.1 Results of Biased Classifier Using p*

The classification results for the biased classifier generated using the P* method for different p-value cases corresponding to different classification cost matrices are shown in Table 3.

7.2 Results of Biased Classifier Using Rebalance

The classification results for the biased classifier generated using the Rebalance method for different p-value cases corresponding to different classification cost matrices are shown in Table 4.

[1] There was high correlation between glucose and insulin variables in the data set. This created ambiguity in the regression model when both were included in the model. Hence the insulin variable was dropped from the regression models.

Table 3. Classification Results of the biased classifier using the p* Method for different p-values.

P-Value	Classification Results				Classification Algorithm Performance Measures			
P*	TP	FP	TN	FN	Accuracy	Specificity	Sensitivity	Classification costs
0.95	2	1	173	79	0.68627	0.99425	0.02469	844.9
0.90	3	1	173	78	0.6902	0.99425	0.03704	844.7
0.85	12	1	173	69	0.72549	0.99425	0.14815	842.84
0.80	15	2	172	66	0.73333	0.98851	0.18519	840.25
0.75	20	5	169	61	0.74118	0.97126	0.24691	834.4
0.70	27	5	169	54	0.76863	0.97126	0.33333	829.66
0.65	32	5	169	49	0.78824	0.97126	0.39506	823.4
0.60	35	7	167	46	0.79216	0.95977	0.4321	815.2
0.55	40	11	163	41	0.79608	0.93678	0.49383	801.8
0.50	49	13	161	32	0.82353	0.92529	0.60494	784
0.45	56	17	157	25	0.83529	0.9023	0.69136	761.8
0.40	59	24	150	22	0.81961	0.86207	0.7284	732.32
0.35	62	30	144	19	0.80784	0.82759	0.76543	703.16
0.30	66	43	131	15	0.77255	0.75287	0.81481	651.2
0.25	74	60	114	7	0.73725	0.65517	0.91358	567
0.20	78	73	101	3	0.70196	0.58046	0.96296	553.6
0.15	79	102	72	2	0.59216	0.41379	0.97531	558
0.10	80	140	34	1	0.44706	0.1954	0.98765	564
0.05	81	166	8	0	0.34902	0.04598	1	569.8
0.01	81	174	0	0	0.31765	0	1	603

7.3 Hypothesis Testing

A T-test was done on the classification cost measures obtained from the 18 classifiers using different p* values to check if there is a significant difference between the mean classification cost of the biased classifiers obtained using p* methods and with the classification cost of the unbiased classifier (784). The results of the T-test shows significant difference ($t = -2.218$, $p = .040$) in the mean classification cost for the biased classifiers at the 95% confidence level ($t_{a=.05, df=18} = 1.96$).

A T-test was also done on the classification cost measures obtained from the 18 classifiers using rebalanced training set samples to check if there is a significant difference between the mean classification cost of the biased classifiers obtained using the rebalance method and with the classification cost of the unbiased classifier (784). The results of the T-test shows significant difference ($t = -2.264$, $p = .035$) in the mean classification cost for the biased classifiers at the 95% confidence level ($t_{a=.05, df=18} = 1.96$).

Table 4. Classification Results of the biased classifier using the Rebalance Method for different p-values

Rebalance Parameters			Classification Results				Classification Algorithm Performance Measures			
Positive cases	Negative cases	Rebalance multiplier	TP	FP	TN	FN	Acc	Specificity	Sensitivity	Class Costs
168	5719	nX19	0	0	174	81	0.682	1.000	0.000	846
168	2709	nX9	3	1	173	78	0.690	0.994	0.037	844.7
168	1703	nX5.6	10	1	173	71	0.718	0.994	0.123	843.2
168	1204	nX4	14	1	173	67	0.733	0.994	0.173	841.5
168	1204	nX4	14	1	173	67	0.733	0.994	0.173	834.73
168	903	nX3	19	5	169	62	0.737	0.971	0.235	829.66
168	702	nX2.3	27	5	169	54	0.769	0.971	0.333	823.4
168	559	nX1.8	32	5	169	49	0.788	0.971	0.395	812.84
168	452	nX1.5	37	8	166	44	0.796	0.954	0.457	797.48
168	368	nX1.2	44	12	162	37	0.808	0.931	0.543	770
168	246	nX0.8	57	19	155	24	0.831	0.891	0.704	751.4
168	201	nX0.6	58	25	149	23	0.812	0.856	0.716	717.24
168	162	nx0.5	63	33	141	18	0.800	0.810	0.778	679.06
168	129	nX0.4	67	45	129	14	0.769	0.741	0.827	622.7
168	100	nX0.3	71	60	114	10	0.725	0.655	0.877	541.7
168	75	nX0.2	78	74	100	3	0.698	0.575	0.963	540
168	53	nX0.1	78	90	84	3	0.635	0.483	0.963	555.5
168	33	nX0.1	79	107	67	2	0.573	0.385	0.975	592.7
168	16	nX0.05	80	133	41	1	0.475	0.236	0.988	672
168	3	nX0.01	80	140	34	1	0.447	0.195	0.988	607

This supports Hypothesis 1 that the expected classification costs of the biased classifiers obtained using the either the p* or the rebalance methods will be significantly different from the classification costs obtained from the unbiased classifier.

7.4 Testing Hypothesis 2

A Paired Samples T-test was done on the accuracy measures of the 36 biased classifiers obtained from the p* method and the rebalance method (18 from each of the methods). The results of the T-test shows no significant difference ($t = -1.386$, $p=.183$) in the mean scores for the predictive accuracy of the biased classifiers from the p* method and the rebalance methods at the 95% confidence level ($t_{a=.05, df=18} = 1.96$).

This supports the hypothesis 2 that the predictive accuracy of the biased classifiers obtained using p* method and the rebalance method will not be significantly different.

7.5 Testing Hypothesis 3

A Paired Samples T-test was done on the classification costs of the 36 biased classifiers obtained from the p* method and the rebalance method (18 from each of the methods). The results of the T-test shows no difference ($t = -597$, $p = .558$) in the mean scores for the classifications costs of the biased classifiers from the p* method and the rebalance methods at the 95% confidence level ($t_{a=.05, df=18} = 1.96$).

This supports the hypothesis that the expected classification costs of the biased classifiers obtained using p* method and the rebalance method will not be significantly different.

7.6 Support for Proposition 1

When P* is very low (less than 0.20) the number of negative training cases in the rebalancing method became very low. Hence, the biased classifier obtained with the rebalance method tended to over-learn the given examples and performed relatively poorly as measured by accuracy on the test data set (when compared to the P* method).

Likewise, when P* was very large (greater than 0.80) the number of positive training cases in the rebalancing method became very low. Hence, the biased classifier obtained with the rebalance method tended to over-learn the given examples and performed relatively poorly as measured by accuracy on the test data set (when compared to the P* method).

8 Discussion of Results

The experimental results indicate support for all of the hypotheses in the study. Therefore, Elkan (2001)'s theorems are supported by our data and experiments. The results of obtaining a biased classifier either by using the P* value or the rebalance method (using the formula in Elkan's paper) is the same in terms of predictive accuracy of the classifier and the expected classification costs of the classifier (Hypotheses 2 and 3).

As expected and predicted, the accuracy and classifications costs do differ for the unbiased classifier and the biased classifier, whether the biased classifier is obtained using the p* method or the rebalance method. As the P* value was increased beyond 0.5, the sensitivity went down while specificity went up. Conversely, as P* was lowered below 0.5, sensitivity increased while specificity decreased (Hypothesis 1).

Our results indicate that in the case of diabetes detection, classification cost matrices need to be skewed so that detection is emphasized. False negatives are almost universally much more expensive than false positives. Clearly, if a person is classified as a false negative and that person is truly diabetic, then the person's health

could significantly deteriorate and ultimately treating the patient will be a lot more expensive. Hence for diabetes data and classification algorithms to detect diabetes – p* values of under 0.2 are appropriate.

It was observed that the P* method works better than the rebalance method as the p-value become smaller and smaller- close to 0.20 and lower. This is because the number of negative examples in the training set used by the rebalancing method becomes extremely low as the p* value is lowered. As a result, the classifier tends to over learn and its performance becomes extremely poor (Proposition 1).

9 Contributions

The following findings are the significant contributions from this research:

- Classification cost structures for Diabetes detection are skewed, as the cost of a false negative is much higher than a false positive.
- The variables PlasmaGlucose and Insulin cannot be used together in a Logistic Regression model as they are highly correlated.
- The valid p* values for diabetes detection data fall in the range of under 0.2.
- The results of using the p* method or rebalance method by adjusting the number of negative training cases gives similar results in terms of accuracy and expected classification costs for the same classification cost matrix.
- Due to the low p* values for the classification cost matrix for diabetes, the rebalancing method should not be used in the diabetes detection classification scenarios as the number of negative training cases becomes very low resulting in over-learning.

References

1. Sagillito, V.: Pima Indians Diabetes Database, http://www.ics.uci.edu/~mlearn/databases/pima-indians-diabetes/ pima-indians-diabetes.names (1990).
2. Ding, B., Gentleman, R.: Classification Using Generalized Partial Least Squares, Collection of Biostatistics Research Archive, http://www.bepress.com/bioconductor/paper5 (2004).
3. Kapat, P., Wang, K.: Classification Using Bayesian Logistic Regression: Diabetes in Pima Indian women Example, http://www.stat.ohio-state.edu/~goel/STAT825/PROJECTS/KapatWang_Team4Report.pdf (2006).
4. Domeniconi, C., Dimitrios, G., Peng, J.: Large Margin Nearest Neighbor Classifiers, IEEE Transactions on Neural Networks, Vol. 16, No. 4, (2005) 899-909.

5. Elkan, C.: The Foundations of Cost Sensitive Learning, Proceedings Seventh International Joint Conference on Artificial Intelligence (IJCAI01), (2001) 973 - 978.
6. Turney, P.: Types of cost in inductive concept learning, Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning (WCSL at ICML-2000), Stanford University, California, (2000) 15-21.
7. Ting, K.M.: A Comparative study of Cost-Sensitive Boosting Algorithms, Proceedings of the Seventeenth International Conference on Machine Learning, June (2000) 983-990.
8. Quinlan, J.R.: C4.5: Programming for machine learning, Proceedings San Mateo: Morgan Kaufmann (1993).
9. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S.: Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, Proceedings of the Symposium on Computer Applications and Medical Care, IEEE Computer Society Press (1988), 261-265.
10. Zadrozny B., Elkan C.: Learning and making decisions when costs and probabilities are both unknown, Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, (2001) 204-212.