



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

(Open Access, Double Blind Peer-reviewed, Refereed and Indexed Journal)

www.iasir.net

A Comparative Analysis of Cost Sensitive Feature Selection Methods on Medical Datasets

Kaushar Ghanchi

Computer Science Department
Navgujarat College of Computer Applications
Ahmedabad, Gujarat, INDIA

Dr. Nileshkumar Modi

Computer Science Department
Baba Saheb Ambedkar Open University
Ahmedabad, Gujarat, INDIA

Abstract: Healthcare mining is one of the most challenging and less explored areas in knowledge discovery. The characteristic of health records are complex nature, missing values and large number of attributes. Some are conditional or dependent and the volume of the data. The most important attribute which may have some associated cost. It is addressed by cost sensitive learning. Costs plays vital role in the selection of feature for further classification. Cost sensitive Feature selection considers different costs while selecting the best performing feature. This research paper focuses on various algorithm of cost sensitive feature selection. We have summarised three methods of cost sensitive feature selection by applying on four dataset of UCI repository. Our study found that on WBCD CASH algorithm performs best but on other datasets cost sensitive decision tree out performs, moreover risk cost used in Cost Sensitive Decision Tree gives better results compare to other algorithms.

Keywords: Cost Sensitive Learning, Feature Selection methods, Misclassification cost, Test cost, Risk cost

I. Introduction

Health Analytics is significant for improved patient care and outcomes. Health records need important analysis to obtain vital information. There are many ways to gain knowledge from health records. All the attributes of the health records may or may not be useful in all types of knowledge extraction. Also medical datasets are having some sorts of unbalancing in class distribution. Imbalanced datasets create several major issues in further process of classification. So most related and important features have to be identified for further process. So it is required to identify the most required and meaningful attribute from the dataset. Attribute selection methods are used to identify the best relevant feature from the dataset. There are many feature selection methods exists like information gain, CFS, CA etc. But these methods are not considering cost while applying function on each feature. Cost plays important role in feature selection on medical dataset. As there are many cost occurs before and after diagnosing a patient. When mining health records cost concept interfere in many key points. A medical practitioner always considers the consequences of misdiagnoses. These costs are non-uniform means failing to detect the disease cost higher than the healthy patient as sick. Another issue is that the tests are very costly and collecting test results delay the time of diagnoses. Though time delayed may not be real cost but need to consider while diagnosing.

Out of all attribute selection methods only cost sensitive learning considers different cost while analysing these data. Health managers and clinicians require some model to minimise several cost associated with health care. We can consider these several types of cost in cost sensitive feature selection and classification process. There are many state of the art algorithm of feature selection exists which consider test cost or misclassification cost or both. Here we focus on these algorithm and apply them on different datasets of medical science.

II. Imbalanced Dataset

Technically any dataset having unequal distribution among its classes is known as imbalanced dataset. It is observed that number of the number of noncancerous patients to exceed greatly the number of cancerous patients. In WBCD, there are 458 benign while 241 malignant cases. Where as in WDBC it is 358 vs 212. And in WPBC, 151 vs 47. Negative i.e benign become majority class while malignant becomes minority class. "Preferably, we require a classifier that provides a balanced degree of predictive accuracy (ideally 100 percent)

for both the minority and majority classes on the data set. In reality, we find that classifiers tend to provide a severely imbalanced degree of accuracy, with the majority class having close to 100 percent accuracy and the minority class having accuracies of 0-10 percent” [1] and it becomes very costly when non-cancerous patients diagnosed as cancerous or vice versa. Unbalancing problem may occur complexities in further mining process. There exist many solutions to solve it. the first solution is sampling methods which is used to reduce the total misclassification cost by using stratification using this methods classifier is not changed but distribution of each class example are modified by under sampling or over sampling . A major issue with these techniques is that it can be used for two class problem also the cost matrix is restricted. Another method is about Meta leaning. They are applicable to base classifier. Many ensemble methods like ADABOOST, ADACOST and METACOST exists in this category. The cost problems are covered by the cost sensitive learning. existing methods involve in alteration of information gain to make it cost sensitive many cost dependent function exists like EG2, EDX and CS-ID3 but very few authors have consider cost other than test or attribute cost and misclassification cost.

Solutions for imbalanced dataset are as follows:

1. Sampling Methods for Imbalanced Learning
 - a. Random Oversampling and Under sampling
 - b. Informed Under sampling
 - c. Synthetic Sampling with Data Generation
 - d. Adaptive Synthetic Sampling
 - e. Sampling with Data Cleaning Techniques
 - f. Cluster-Based Sampling Method
 - g. Integration of Sampling and Boosting
2. Cost-Sensitive Methods\for Imbalanced Learning
 - a. Cost-Sensitive Learning Framework
 - b. Cost-Sensitive Data space Weighting with Adaptive Boosting
 - c. Cost-Sensitive Decision Trees
 - d. Cost-Sensitive Neural Networks
3. Kernel-Based Methods and Active Learning
 - a. Methods for Imbalanced Learning
 - b. Kernel-Based Learning Framework
 - c. Integration of Kernel Methods with Sampling Methods
 - d. Kernel Modification Methods for Imbalanced Learning
 - e. Active Learning Methods for Imbalanced Learning

III. Cost Sensitive Learning

Sampling methods are used to balance the distribution evaluating respective proportion of class example in distribution whereas cost sensitive learning methods consider different costs associated with misclassified examples.[1]. In these methods, cost matrix is used to identify the cost for misclassified examples. A cost sensitive algorithm is very contemporary issue in machine learning research. Cost sensitive learning is superior to sampling methods. Cost sensitive algorithms focus on binary classification problems. There are two classes as positive class i.e. the minority, and the negative class as the majority.

Let $C(i; j)$ be the cost of predicting an example belonging to class i when in fact it belongs to class j ; the cost matrix will be as follows.

		Predicted class	
		Positive	Negative
Actual class	Positive	$C(\text{True positive})$	$C(\text{False Negative})$
	Negative	$C(\text{False Positive})$	$C(\text{True Negative})$

Cost of each cell of confusion matrix is a cost matrix. The cost matrix will provide the costs associated with the four outcomes shown in the confusion matrix, which we refer to as $C(\text{True positive})$, $C(\text{False Negative})$, $C(\text{False Positive})$, and $C(\text{True Negative})$. As is often the case in cost-sensitive learning, we assign no costs to correct classifications, so $C(\text{True positive})$ and $C(\text{True Negative})$ are set to 0. Since the positive (minority) class is often more interesting than the negative (majority) class, typically $C(\text{False Negative}) > C(\text{False Positive})$ (note that a false negative means that a positive example was misclassified). The main objective of cost sensitive learning is to minimize the total cost. Total cost can be found as

Total cost = False Negative * $C(\text{False Negative})$ + False Positive * $C(\text{False Positive})$

IV. Cost Sensitive Feature Selection Methods

There are many cost sensitive feature selection methods exists. Out of all some methods have performs better on medical datasets. Some methods consider only test cost while some methods consider both test cost and misclassification cost. Only one methods considers Risk cost and no one considers delay or any other cost. Following are the cost sensitive feature selection methods which will be tested on various datasets.

1. Inexpensive Classification with Expensive Tests -ICET

Inexpensive Classification with Expensive Tests – ICET is cost sensitive algorithm developed by Peter Turney. This algorithm combines decision tree with Genetic algorithm. At first, this algorithm generates random set of initial population where each individual relate to one decision tree. Then there is a fitness value of each individual computed and evaluated. Then new individual are evolved by mutation and cross over methods. This algorithm computing ICF that is informational cost function. As attribute selection function ICF is computed by applying following cost function.

$$ICF = \frac{2 * I \Delta}{(Ci+1)^\omega}$$

2. Cost Sensitive Decision Tree

Alberto Freitas et al. developed cost sensitive decision tree methods by considering risk cost. Authors have given the cost function for decision tree spiting criteria to contemplate cost function CF.

$$\text{Cost function is } \frac{I \Delta i}{(Ci \phi i)^\tau}$$

Where I is the information gain, C is attribute cost and Q is risk cost associated with attribute. Here author also considers l cost scale factor. Here authors have used concept of risk to penalize attribute that might be invasive or may cause discomfort and disturb the quality of life of patient. The value of risk cost 1 means absence of risk while higher value of risk cost means it is harmful to the patient.

3. Cost Sensitive attribute selection using Histogram

Yael Weiss et al. had developed the CASH algorithm that is cost sensitive attribute selection using histogram. Author calculates a priori cost of dataset and then divide the dataset according to the best split of the concern feature. Here algorithm identifies the misclassification cost ratio of the attribute to calculate average total cost of the same attribute. If the total average cost of the attribute is smaller than the average a priori cost then this attribute will be inserted in the initial population for further classification i. e the attribute will be selection. Here author uses misclassification cost and test or attribute cost for the calculation of a priori cost and average total cost. Authors have used genetic algorithm's operator mutation and crossover for selection process.

V. Experiment Results

We have used four data sets from UCI repository for our experiment. Following table shows the detail of the dataset and its corresponding class. All four datasets are medical dataset where PIMA shows the information about the diabetes patient detail. HIPATITIS is about the liver disease patient data. We have used Wisconsin breast cancer dataset and BUPA liver disorder dataset for the experiment purpose. All these four datasets have imbalance in class distribution. Also the values in the attributes are real or integer. Some fields are having values in some frequency. Following table shows the detail of all four datasets.

	Pima	Hipatitis	Wbcd	Bupa
Description	Pima Indians Diabetes Database	Hepatitis Domain	Wisconsin Breast Cancer Dataset	BUPA liver disorders
Total Records	768	155	699	345
Total Attributes	9	20	11	7
Class Distribution	CLASS 0 : 500	DIE:32	Benign : 458	Class 1: 145
	CLASS 1 : 268	LIVE:155	Malignant :241	Class 2: 200

We have applied three cost sensitive methods for feature selection on above four datasets. We took various misclassification cost and test cost values for the evaluation of the algorithm.

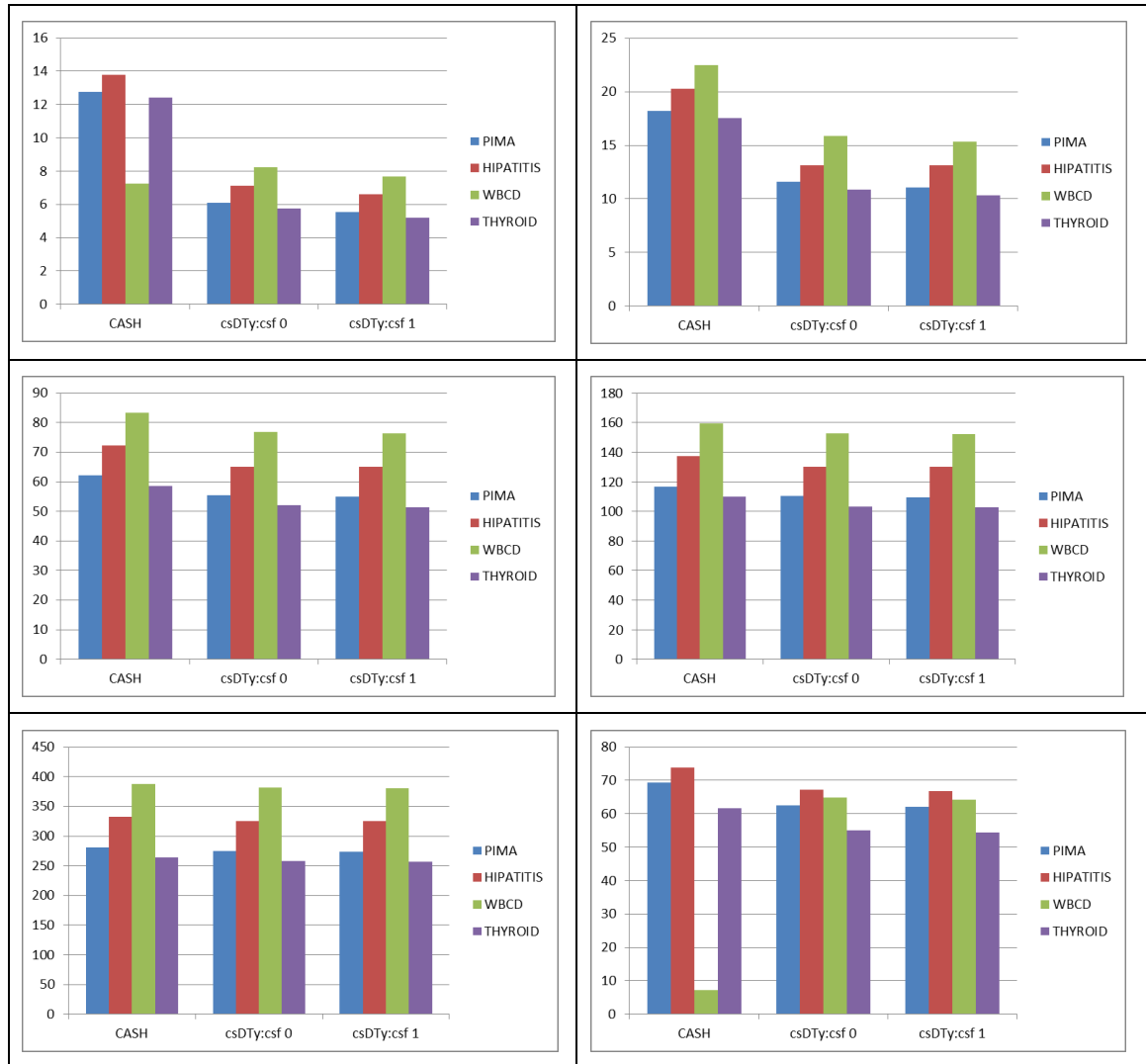
Following figure shows the calculation of MCR, Negative and positive Misclassification cost and Average Total Cost of Hepatitis Dataset. Here we take False Positive and False Negative ratio as 10/10. Same way we have calculated values for all other datasets.

	CLASS DIE - POSITIVE		CLASS LIVE - NEGATIVE	
SPLIT	<=1	>1	<=1	>1
TOTAL RECORDS IN EACH SPLIT	9	23	87	19
TO MULTIPLY WITH FP/FN	10	10	10	10
	90	230	870	190
POSITIVE/NEGATIVE MC	0.09375	0.54762	0.90625	0.45238
TOTAL MCR	8.4375	125.952	788.438	85.9524
TOTAL MCR				1008.78
AVERAGE MCR				6.50826
AVERAGE TOTAL COST				2.75565

We have calculated ATC for different combinations of FP/FN values. We took combinations of 10/10,20/20,100/100,200/200,500/500. These are the same values of False Positive and False Negative. Then we took different values for both FP and FN as 100/10,10/100,100/200 etc. following table shows the corresponding values of ATC for all these runs of FP and FN. Here we found that CSDT with cost scale factor 1 outperforms with the lowest Average Total Cost in almost all combinations of FP and FN. CASH algorithm outperforms in some of the dataset with FP/FN combination of 10/10. Moreover we observed that the best ATC generated is 5.22011 on THYROID dataset by CSDT with cost scale factor 1 while the highest ATC is 994.4979 on WBCD dataset by CASH method.

Figure 1 . Average total cost for different values of FP and FN

DATASET	ALGORITHM	MISCLASSIFICATION COST MATRIX														
		10	20	100	200	500	10	100	10	100	200	100	500	1500	500	
		10	20	100	200	500	100	10	20	200	100	500	100	500	1500	
PIMA	CASH	12.74882	18.22764	62.05821	116.8464	281.2111	69.24666	37.73448	18.29642	117.5342	75.16638	304.7732	152.4372	456.8895	863.084	
	csDTy:csf 0	6.087821	11.56664	55.39721	110.1854	274.5501	62.58566	31.07348	11.63542	110.8732	68.50538	298.1122	145.7762	450.2285	856.423	
	csDTy:csf 1	5.56259	11.04141	54.87198	109.6602	274.0248	62.06043	30.54825	11.11019	110.348	67.98015	297.587	145.251	449.7033	855.8977	
HIPATITIS	CASH	13.77826	20.28651	72.35257	137.4351	332.6828	73.92493	26.78129	20.059	135.16	80.3533	334.5803	119.5731	429.8314	974.4474	
	csDTy:csf 0	7.117257	13.10028	65.16633	130.2489	325.4966	67.26393	20.12029	12.87277	127.9738	73.16707	327.3941	112.3869	422.6451	967.2612	
	csDTy:csf 1	6.592025	13.10028	65.16633	130.2489	325.4966	66.7387	19.59506	12.87277	127.9738	73.16707	327.3941	112.3869	422.6451	967.2612	
WBCD	CASH	7.270044	22.4992	83.41599	159.562	387.9999	7.272191	39.76796	20.86376	143.2076	103.8314	329.9847	181.1657	608.881	994.4979	
	csDTy:csf 0	8.223599	15.8382	76.75499	152.901	381.3389	64.79782	33.10696	14.20276	136.5466	97.17044	323.3237	174.5047	602.22	987.8369	
	csDTy:csf 1	7.698368	15.31297	76.22976	152.3757	380.8137	64.27259	32.58173	13.67752	136.0213	96.6452	322.7985	173.9795	601.6948	987.3116	
THYROID	CASH	12.40635	17.54269	58.63347	109.9969	264.0874	61.61	44.91638	16.9069	103.639	80.81325	265.8017	184.64	531.1337	748.8923	
	csDTy:csf 0	5.745347	10.88169	51.97247	103.3359	257.4264	54.949	38.25538	10.2459	96.97801	74.15225	259.1407	177.979	524.4727	742.2313	
	csDTy:csf 1	5.220116	10.35646	51.44724	102.8107	256.9011	54.42377	37.73015	9.72067	96.45278	73.62702	177.4538	177.4538	523.9475	741.706	



VI. Conclusion

We have used four medical dataset for experiment of cost sensitive feature selection methods CASH,CSDT with cost scale factor 0 and 1. This study is on various runs of several values of False positive and False Negative. The results shows that the cost sensitive decision tree with cost scale factor 1 out performs in almost all runs. While one breast cancer dataset CASH gives better results than CSDT. We observe that cost scale factor and risk cost plays vital role in the selection process. This study can be extended in development of novel architecture of cost sensitive feature selection by considering risk cost and delayed cost along with the misclassification and test or attribute cost. With regard to the future research, several possibilities for improvement of results came in our mind. One of the important thing is to consider several other cost associated with the dataset. It might be separate cost of all datasets. Another important possibility is to deploy other better methods for the spiting criteria than histogram. We can think of chi merge or entropy for identifying the best split for the process.

VII. References

- [1] Nileshkumar Modi and Kaushar Ghanchi,"A Comparative Analysis of Feature Selection Methods and Associated Machine Learning Algorithms on Wisconsin Breast Cancer Dataset (WBCD)",Journal on Advances in intelligent Systems and Computing,ICT4SD 2015 Volume 1
- [2] Haibo He, Member, IEEE, and Eduardo A. Garcia,"Learning from Imbalanced Data",Ieee Transactions On Knowledge and Data Engineering, Vol. 21, No. 9, September 2009
- [3] Charles X. Ling, Victor S. Sheng,"Cost-Sensitive Learning and the Class Imbalance Problem", Encyclopedia of Machine Learning. C. Sammut (Ed.). Springer. 2008
- [4] K.P.N.V .Satyasree, Dr.J.V.R.Murthy, "An Exhaustive Literature Review on Class Imbalance Problem",International Journal of Emerging Trends" Volume 2, Issue 3, May – June 2013
- [5] P. Turney, Types of cost in inductive concept learning, Proc. Workshop Cost-Sensitive Learning, in: 17th International Conference on Machine Learning, 2000, pp. 15–21.

- [6] Freitas, A. da Costa-Pereira, P. Brazdil, Cost-sensitive decision trees applied to medical data, in: DaWak, Lecture Notes in Computer Science, 2007, pp.303–312.
- [7] Hong hao, Fan Min, and William Zhu," Cost-Sensitive Feature Selection of Numeric Data with Measurement Errors",Hindawi Publishing Corporation, Journal of Applied Mathematics Volume 2013, Article ID 754698, 13 pages <http://dx.doi.org/10.1155/2013/754698>
- [8] Yael Weiss, Yug val Elovici, Lior Rokach,"The CASH algorithm-cost-sensitive attribute selection using histograms",Inform. Sci.(2011), doi:10.1016/j.ins.2011.01.035
- [9] Divya Tomar and Sonali Agarwal,"Hybrid Feature Selection Based Weighted Least Squares Twin Support Vector Machine Approach for Diagnosing Breast Cancer, Hepatitis, and Diabetes",Hindawi Publishing Corporation Advances in Artificial Neural Systems, Volume 2015, Article ID 265637, 10 pages, <http://dx.doi.org/10.1155/2015/265637>
- [10] X. Chai, L. Deng, Q. Yang, C.X. Ling, Test-cost sensitive Naive Bayes classification, in: Proceedings of 4th International Conference on Data Mining, 2004, pp. 51–58.
- [11] Fang, J.R. Rajasekera, and H.S. J. Tsao, Entropy Optimization and Mathematical Programming, Kluwer Academic Publishers, Norwell, 1997.
- [12] D.Lavanya and Dr.K.Usha Rani,"ENSEMBLE DECISION TREE CLASSIFIER FOR BREAST CANCER DATA",International Journal of Information Technology Convergence and Services (IJITCS) Vol.2, No.1, February 2012
- [13] Yuwen Huang,"Research on cost-sensitive ensemble classification for mining imbalanced massive data streams",COMPUTER MODELLING & NEW TECHNOLOGIES 2014 18(11) 375-382
- [14] Nitesh V. Chawla,DATA MINING FOR IMBALANCED DATASETS:AN OVERVIEW,chapter
- [15] Petr Berka,Jan Rauch,Djamel Abdelkader Zighed,"Data Mining and Medical Knowledge Management: Cases and Applications" BOOKPublished in the United States of America by Information Science Reference (an imprint of IGI Global)
- [16] Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

VIII. Acknowledgements

Wisconsin Breast Cancer dataset (WBCD), Wisconsin Diagnostic Breast Cancer (WDBC) and Wisconsin Prognostic Breast Cancer (WPBC) databases were obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. We have used these datasets for further research and thankful for providing datasets.