ProICET: Case Study on Prostate Cancer Data*

Camelia Vidrighin, Rodica Potolea, Ioana Giurgiu, Mihai Cuibus

Technical University of Cluj-Napoca, 26 Baritiu St, Cluj-Napoca, Romania
Camelia.Vidrighin@cs.utcluj.ro, Rodica.Potolea@cs.utcluj.ro

Cancer is the second most threatening disease in the world today, not only because of its mortality rate, but also due to the brutal changes it imposes on the patient's life, and the fact that its exact causes of progression remain to be discovered. Prostate cancer is the second cause of cancer-related deaths in men in the USA – representing a hazard especially for African-American men over the age of 50. Recent evolution in computer technology has resulted in the emergence of a combined approach to the diagnosis and prognosis process, with a data driven analytical approach complementing biomedical and clinical methods. Cost sensitive learning is one such data mining method, particularly well suited for medical problems. This paper investigates the performance of a new system based on a hybrid cost-sensitive algorithm (*ProICET*) on a prostate cancer medical dataset, while trying to produce new medical knowledge.

Keywords
Cost-sensitive learning, data mining, hybrid algorithm, implementation, prostate cancer

1. Introduction

Recent statistics show that cancer has become one of the most serious threats to human life nowadays, being the second cause of death in the USA in 2004. The American Cancer Society estimates that prostate cancer in particular will be responsible for 9% of the total cancer-related deaths in men in 2007 [1].

A cancer diagnosis has obviously a huge impact on human life, affecting especially the patient's emotional state, but also his/her lifestyle and life expectations. Therefore it is natural that patients wish to know their prognosis and understand what they are dealing with in terms of treatment, quality of life, or finances. Research in the biological and clinical methods for cancer diagnosis and prognosis has lead to a better understanding of the likely course and outcome of this disease. Still, despite the boost in biomedical technology, the accuracy of diagnosis and prognosis is, in many cases, rather low, mainly because of the negative influence of factors like the physician's experience, intuition and biases, or the large amount of data to be analyzed.

In this context, machine learning can be used to automatically infer diagnostic rules from descriptions of past, successfully treated patients, and help specialists make the diagnostic process more objective and more reliable. Records of previous patients are gathered into hospital archives and can be made available through machine learning techniques; the classifier derived from this data provides support for future diagnosis and treatment and can help improve the physician's speed, accuracy and reliability in establishing a new diagnosis.

Moreover, thanks to the large variety of available machine learning mechanisms, and the possibility to store and process large amounts of data, new associations and interdependencies can be discovered, leading to new medical research directions and suggestions for new approaches to the diagnosis process.

* This work was supported by the grant no. 18CEEX-I03/2005, founded by the Romanian Ministry for Education and Research

1.1. Prostate Cancer

Prostate cancer occurs when cells of the prostate (gland in the male reproductive system) mutate and begin to multiply out of control, spreading to other parts of the body (bones and lymph nodes mainly). Symptoms include pain, difficulty in urinating, erectile dysfunction, and many others.

Physical examination and PSA (Prostate Specific Antigen) blood tests are crucial for early diagnosis of the disease. Confirmation is received upon performing a biopsy of the prostate tissue. Further investigations, such as X-rays and bone scans, may be performed to determine the degree of spread.

There are several possible treatments for prostate cancer, such as: surgery, radiation therapy, chemotherapy, hormone therapy, or a combination of these – depending on the extent of spread of the disease, age and general state of health, and so on.

Just like with any other cancer type, the exact causes that trigger the metastasis remain unidentified. There are still some factors that are known to influence the evolution of the disease, such as: age (the risk increases quickly after the age of 60), race (African-American people are more susceptible to developing metastasis), genetics and diet (a diet rich in animal fats is very dangerous).

1.2. Medical Data Mining

Medical data mining is considered to be one of the most challenging areas of application in knowledge discovery. Main difficulties are related to the complex nature of data (heterogeneous, hierarchical, time series), or to its quality (possibly many missing values) and quantity. Domain knowledge or ethical and social issues are also of great importance. [2]

But maybe the most important particularity of medical data mining problems is the concept of *cost*, which is addressed by cost-sensitive classification (to be discussed in the next section).

When mining a medical problem, the concept of cost interferes in several key points. First of all, a doctor must always consider the potential consequences of a misdiagnosis. In this field, misclassification costs may not have a direct monetary quantification, but they represent a more general measure of the impact each particular misclassification may have on human life. These costs are non-uniform (diagnosing a sick patient as healthy carries a higher cost than diagnosing a healthy patient as sick). Another particularity of the medical diagnosis problem is that medical tests are usually costly. Moreover, collecting test results may be time-consuming. Arguably, time may not be a real cost, but it does have some implication for the decision whether it is practical to take a certain test or not. In the real case, performing all possible tests in advance is unfeasible and only a relevant subset should be selected. The decision on performing or not a certain test should be based on the relation between its cost and potential benefits. When the cost of a test exceeds the penalty for a misclassification, further testing is no longer economically justified.

1.3. Cost Sensitive Learning

Recent studies [3], [4] suggest that more complex measures for evaluating the adequacy of classifiers should be considered, rather than the classical error reduction strategy. For example, non-uniform costs are inherent in fields such as medical diagnosis (classifying a healthy patient as ill is far less expensive than the reverse situation), or fraud detection. Such problems are addressed by *cost-sensitive classification*, which is directed towards the reduction of the total cost, instead of just minimizing the number of misclassification errors.

Turney [3] provides a general taxonomy of costs involved in inductive concept learning, the most important of which being *misclassification costs* and *test costs*. The first category encompasses the costs which are conventionally considered by most cost-sensitive classifiers; however, several solutions address the second category also. A brief survey of

the most important cost-sensitive classifiers, as described in the literature, is provided in the following.

Chronologically one of the first solutions for the problem of reducing the total misclassification cost was a procedure called *stratification*. In this approach, the actual classifier is not altered in any way; instead, the distribution of examples for each class is changed, either by undersampling or by oversampling. The main drawback of this technique is that it restricts the form and dimension of the cost matrix (it is only appropriate for two-class problems or problems where the cost is independent of the predicted class).

More complex techniques, which overcome these limitations, usually involve meta-learning algorithms that are typically applicable to a range of base classifiers. In this category we include algorithms based on various *ensemble* methods, such as *AdaBoost.M1* [5], *AdaCost* [6], or *MetaCost* [7].

Several approaches exist also for tackling the problem of test costs. They usually involve some alteration of the information gain function, as to make it cost-sensitive. Various cost dependent functions have been proposed in the literature, such as EG2, IDX or CS-ID3 [4].

Significantly less work has been done for aggregating both cost components. The most prominent approach in the literature is ICET, which combines a greedy search heuristic (decision tree) with a genetic search algorithm. Other possible solutions are explored in [8], [9] and [10].

As shown in section 1.2, medical diagnosis is one field in which such an aggregated approach is of utmost importance.

2. The Basic ICET Algorithm

As most greedy techniques, classical tree induction suffers from the horizon effect (it tends to get caught at local optima). One possibility to avoid this pitfall is to perform a heuristic search in the space of possible decision trees through evolutionary mechanisms.

One of the most prominent algorithms that explore this idea is ICET (Inexpensive Classification with Expensive Tests). Introduced by Peter Turney, the technique tackles the problem of cost-sensitive classification by combining a greedy search heuristic (decision tree) with a genetic algorithm [4].

ICET begins with the genetic algorithm (GA) generating randomly a set of individuals (the initial population), each individual corresponding to one decision tree. The fitness value of each individual is evaluated in the decision tree component (to be described shortly). Then, for a specific number of iterations, new individuals are evolved, by applying standard mutation and crossover operators, and their fitness is computed. After the last iteration, the fittest individual is returned – its biases are used to train the output classifier.

The genetic algorithm used in [4] is GENESIS, and the decision tree algorithm is a modified version of Quinlan's C4.5 [11], which uses ICF (Information Cost Function) as attribute selection function, same as in EG2. For the i^{th} attribute, ICF may be defined as follows:

$$ICF_i = \frac{2^{\Delta I_i}}{(C_i + 1)^w}, \quad where \quad 0 \le w \le 1 \tag{1}$$

The formula above shows that, when building a decision tree, the attribute selection criterion is no longer based solely on the attribute's contribution to obtaining a pure split (the information gain ΔI_i), but also on its cost, C_i. For example, when choosing between two attributes which have the same value for the information gain, the attribute with the lower cost will be preferred. Parameter w adjusts the strength of the bias towards lower cost attributes. When $w = 0$, the cost of the attribute is ignored, and selection by ICF is equivalent to selection by the information gain function; when $w = 1$, ICF is strongly biased by the cost component.

An important remark is that, unlike EG2, ICET does not minimize test costs directly. Instead, it uses ICF for the codification of the individuals in the population. The n costs, C_i, are not

true costs, but *bias parameters*. They provide enough variation to prevent the decision tree learner from getting trapped in a local optimum, by overrating/ underrating the cost of certain tests based on past trials' performance. However, it is possible to use true costs, when generating the initial population, which has been shown to lead to some increase in performance.

Each individual (decision tree) is represented as a Gray encoded bit string, corresponding to $n + 2$ numbers. The first n numbers are the bias parameters ('alleged' test costs in the ICF function). The last two stand for the algorithm's parameters *CF* and *w*; the first controls the level of pruning (as defined for C4.5), while *w* is needed by ICF.

The *fitness function* for an individual is computed as the average cost of classification of the corresponding tree (obtained by randomly dividing the training set in two subsets, the first used for the actual tree induction and the second for error estimation). The average cost of classification is obtained by normalizing the total costs (obtained by summing the test and misclassification costs) to the training set size.

Test costs are specified as attribute - cost value pairs. The classification costs are defined by a cost matrix $(C_{ij})_{nxn}$, where C_{ij} is the cost of misclassifying an instance of class *j* as being of class *i*. If the same attribute is tested twice along the path (numeric attribute), the second time its cost is 0.

2.1. ProICET – New System Based on ICET

We have implemented the theoretical model presented above as a new system – *ProICET*. As a starting point we used the implementation of the C4.5 algorithm, revision 8, provided by Weka (referred to as J4.8), and a general-purpose genetic algorithms library called GGAT (General Genetic Algorithm Tool*)*, both written in Java.

Some enhancements have been considered in the genetic component, mainly in what the genetic parameters are concerned. For each individual, the $n + 2$ chromosomes have been defined (*n* being the number of attributes in the data set, while the other two correspond to parameters *w* and *CF*); each chromosome is represented as a 14 bits binary string, encoded in Gray. The population size is 50 individuals. The *roulette wheel* technique was used for parent selection; as recombination techniques, we have employed *single point random mutation* with mutation rate 0.2, and *multipoint crossover*, with 4 randomly select crossover points.

The most important changes are the use of *elitism* and the *single population* technique, which allow exceptional individuals to propagate unaltered to future generations. Also, we used the *fitness ranking* method to compare the individuals' strengths, in order to avoid the situation when only a few elements, which are by far stronger than the rest, have very high probability of being used as parents (thus reducing the search variability).

The number of evaluation steps has also been increased to 1000. Due to the fact that a new generation is evolved using single population, the final result yielded by the procedure is the best individual over the entire run, which makes the decision on when to stop the evolution less critical. More than that, experiments show that usually the best individual does not change significantly after 800 steps.

3. Experiments

In [12], the objective was to validate the soundness of *ProICET* when compared to several other well known cost-sensitive algorithms (MetaCost, EG2), as well as the best classical decision tree learner (C4.5) and AdaBoost. Therefore, a comparative analysis of the misclassification cost component was provided on several benchmark medical datasets, and also an analysis of the behaviour of the algorithm in real-world conditions, with both test and misclassification costs. In evaluating the results we observed that *ProICET* yielded lower costs than the other algorithms. Also, it achieved very high accuracy rates on reasonably

sized to large datasets (94% on Wisconsin Breast Cancer, consisting of 699 instances, 99% on Thyroid, 7200 instances).

Hence, the strategy of enhancing decision tree induction with evolutionary means has proved to be beneficial. This is shown by the comparison of the cost values obtained with *ProICET* and those with EG2 alone (with no genetic component), as well as by the relative comparison with other well-known systems [12].

The objective of this paper is to evaluate the system on a real prostate cancer dataset. The main goals are to verify that it maintains its behaviour in this real-world medical problem, yielding low costs while maintaining a high precision rate. A second direction of investigation involves ranking the predictor attributes, such as to try and match results obtained by our system with medical staff assumptions.

The dataset was provided by the Medicine and Pharmacy University of Cluj-Napoca. During the discussions with the medical team, a set of major/immediate interest parameters were defined. Consequently, data cleaning was performed and the final version of the dataset for the current investigation stage was obtained.The attributes employed can be seen in Table 1.

Since the algorithm involves a large heuristic component, the evaluation procedure assumes averaging the costs over 10 runs. Each run uses a pair of randomly generated training-testing sets, in the proportion 70% - 30%; the same proportion is used when separating the training set into a component used for training and one for evaluating each individual (in the fitness function). We used two different values for the test costs – 0 and 0.1 – and four different cost matrices (built such as to emphasize the unbalance in different errors' severity). This resulted in eight different batches.

Table 1 – Prostate cancer dataset

Attribute	**Range**
One (*TNM*)	Symbolic (1a, 1b, 1c, 2a, 2b, 3a, 3b)
Two (*Gleason Score*)	Numeric (2-10)
Three (*Presence on Median Intravesical Lobe*)	Symbolic (not present in the ultrasound, voluminous, intravesical)
Four (*Prostate Volume*)	Numeric
Five (*Preoperative PSA*)	Numeric (ng/ml)
Six (*IIEF - International Index of Erectile Function*)	Numeric
Seven (*Quality of Life*)	Numeric (0-2)
Eight (*Surgery Type*)	Symbolic (TP, EP)
Nine (*Operative Technique*)	Symbolic (Ante Grade, Retro Grade, Bipolar)
Ten (*Nerve Sparing*)	Symbolic (Non, NS Left, NS Right)
Eleven (*Bleeding*)	Numeric (minutes)
Twelve (*Anastomosis*)	Symbolic (Continuous, Separate, Van Velt)
Thirteen (*Operative Time*)	Numeric (minutes)
Fourteen (*Postoperative Hospitalization*)	Numeric (days)
Fifteen (*Complications*)	Boolean
Class (Postoperative PSA)	Symbolic (low: PSA < 0.1, medium: PSA $\in$ [0.1, 1], high: PSA > 1)

The cost matrices are shown in tables 2-5 (AC – Actual Class, PC – Predicted Class). The main idea in building them was to capture the different cost of errors as well as possible, while keeping a reasonable ratio between them.

The results for the eight different batches are presented in Table 6. We observe that, when both types of costs are considered, *ProICET* yields the lowest total costs, which proves once again it is the best approach for cost reduction in medical problems.

Table 2 – Cost matrix 1

AC \ PC	low	medium	High
low	0.0	0.5	1.0
medium	1.5	0.0	0.7
high	5.0	3.0	0.0

Table 3 – Cost matrix 2

AC \ PC	low	medium	High
low	0.0	0.5	1.0
medium	3.0	0.0	0.7
high	10.0	6.0	0.0

Table 4 – Cost matrix 3

AC \ PC	low	medium	high
low	0.0	0.5	1.0
medium	0.75	0.0	0.7
high	2.5	1.5	0.0

Table 5 – Cost matrix 4

AC \ PC	low	medium	high
low	0.0	0.5	1.0
medium	3.0	0.0	0.5
High	5.0	3.0	0.0

Table 6 – Total costs and accuracy rate
(TC – value of Test Costs; CM – Cost Matrix)

Medical Dataset	**Average Accuracy Rate**					**Average Total Cost**				
	Pro ICET	**Ada Boost**	**EG2**	**J4.8**	**Meta Cost**	***Pro ICET***	**Ada Boost**	**EG2**	**J4.8**	**Meta Cost**
TC:0,TM:1	84.18%	79.18%	84.07%	84.07%	84.18%	0.28	0.284	0.269	0.269	0.293
TC:0.1,TM:1	83.77%					**0.414**	0.734	0.430	0.430	0.448
TC:0,TM:2	83.87%				83.26%	0.561	0.52	0.52	0.52	0.65
TC:0.1,TM:2	84.07%					**0.678**	0.97	0.682	0.682	0.812
TC:0,TM:3	84.28%				84.38%	0.146	0.166	0.142	0.142	0.145
TC:0.1,TM:3	84.07%					**0.252**	0.616	0.305	0.305	0.310
TC:0,TM:4	84.07%				83.36%	0.213	0.44	0.44	0.44	0.502
TC:0.1,TM:4	83.77%					**0.575**	0.89	0.603	0.603	0.647

The fact that the accuracy rates (~84%) do not reach very high values could be rooted in the characteristics of the dataset: the number of instances was reduced during the pre-processing stage, because of the high number of missing values. We estimate that on a larger dataset the precision rate will be higher. Another piece of evidence that supports our assumption can be found in [12] – where *ProICET* was evaluated against four renowned classifiers (AdaBoost, EG2, J4.8 and MetaCost) and yielded both better costs and better accuracy rates.

Another important result is related to the ranking of the attributes in the order of their prediction power. Since during the training process equal test costs were assigned to each attribute, the cost component did not influence in any way the choice of one attribute over another (it only affects the total cost of the trees in the sense that bigger trees yield higher total costs). By analyzing the output trees we came up with the following list of best predictor attributes (the same list was obtained by the other algorithms as well):

- Four (Prostate Volume)
- Nine (Operation Technique)
- Eleven (Bleeding)
- Two (Gleason Score)
- Six (IIEF)
- Five (Pre-Op PSA)

The fact that the prostate volume appears the first in most tests is new, and, according to the medical team's opinion, it is the confirmation of a fact they have been suspecting for some time now.

4. Conclusions and Future Work

The idea of combining biomedical and clinical expertise with data driven analysis in the medical diagnosis process has gained considerable interest recently. One of the main fields where this approach has proven to be beneficial is that of cancer diagnosis and prognosis, where a cost-sensitive tactic seems to be the most appropriate.

ICET provides such an approach, and previous work has shown it is a robust algorithm [12]. This paper tries to evaluate a new system, *ProICET*, based on the ICET algorithm on a particular prostate cancer dataset, and to provide an answer to a medical question related to the most important attributes that influence the value of postoperative PSA. The results obtained show that *ProICET* is the best at cost reduction when both types of costs are involved. Moreover, we estimate that on larger datasets the accuracy rate will reach higher values than in the current evaluation. This assumption is sustained by previous evaluations on larger datasets, which have shown that normal accuracy rates are in the range 94 – 99%. These values are better than, or comparable with those of algorithms accepted by the machine learning community as being powerful schemes for error/cost reduction ([4], [5], [7]).

The attribute ranking we obtained was validated by the results obtained by other decision tree learners, and was confirmed by a medical specialist. This result is of particular interest to the medical team, because, in addition to validating their suppositions on the importance of the prostate volume in the value of postoperative PSA, it opens new possibilities of expertise in the medical field. Also, starting from the obtained ranking, we can carry out more tests that will show for sure that the ranking obtained is not pure chance (we intend to implement the mutual information measure and evaluate pairs of attributes). Therefore, a study on the attribute correlations could bring about new medical knowledge that is otherwise very difficult to fetch from the huge amount of data.

Since ICET has a strong evolutionary component, one possible way of improving the results is to experiment on the setting of the genetic parameters, such as mutation rate, population size, or crossover method. Also, increasing the size of the training dataset is necessary, in order to obtain a more accurate classifier.

References

[1] American Cancer Society. Cancer Facts&Figures 2007. http://www.cancer.org/docroot/stt/stt_0.asp Accessed in March 2007.

[2] Dursun H, Nainish P. Knowledge extraction from prostate cancer data. Proceedings of the 39th Hawaii International Conference on System Sciences, 2006.

[3] Turney P. Types of cost in inductive concept learning. Proceedings of the Workshop on Cost-Sensitive Learning, 7th International Conference on Machine Learning, 2000.

[4] Turney P. Cost sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. Journal of Artificial Intelligence Research, (2):369–409, 1995.

[5] Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119–139, 1997.

[6] Fan W, Stolfo S, Zhang J, Chan P. Adacost: Misclassification cost-sensitive boosting. Proceedings of the 16th International Conference on Machine Learning, pages 97–105, 2000.

[7] Domingos P. Metacost: A general method for making classifiers cost-sensitive. Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, 1991.

[8] Li J, Li X, Yao X. Cost-sensitive classification with genetic programming. Proceedings of the 2005 Congress on Evolutionary Computation, 3:2114–2121, 2005.

[9] Sheng S, Ling C. Hybrid cost-sensitive decision tree. PKDD, pages 274–284, 2005.

[10] Sheng S, Ling C, Yang Q. Simple test strategies for cost-sensitive decision trees. ECML, pages 365–376, 2005.

[11] Quinlan J. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[12] Vidrighin C, Savin C, Potolea R. A Hybrid Algorithm for Medical Diagnosis. Submitted to the IEEE Region 8 Eurocon 2007 Conference, February 2007.

[13] R. Potolea, C. Vidrighin, C. Savin. ProICET – A Cost-Sensitive System for the Medical Domain. Accepted for ICNC'07-FSKD'07, March 2007.