New Complex Approaches for Mining Medical Data

Camelia Vidrighin Bratu, Rodica Potolea
Technical University of Cluj-Napoca
26-28 Baritiu St., Cluj-Napoca
Camelia.Vidrighin@cs.utcluj.ro

Abstract

The medical field has recently become one of the most challenging application areas for data mining techniques. The particularities involved in mining medical problems, such as domain knowledge, ethical and social issues, data quality, complexity and quantity, or cost, have lead to the necessity for more complex approaches. This paper tries to tackle some of the main issues, by introducing three new systems. ProICET adopts a hybrid approach to reduce the costs involved in the diagnosis process and help avoid "dangerous" errors. A second system, based on a PANE method, is introduced to reduce the misdiagnoses while keeping the transparency of the decision process. Finally, a system which combines different classifier predictions is presented. The benefits of such an approach include the possibility of establishing the baseline accuracy for any dataset, and the capability to consider data coming from different sources, with different structures.

1 Introduction

The medical domain is considered to be one of the most challenging areas of application in knowledge discovery. The main difficulties are related to the complex nature of the data involved (heterogeneous, hierarchical, time series), its quality (possibly many missing values) and quantity. Although hospitals hold huge amounts of records belonging to past treated patients, part of this data is not in electronic form, and the idea of transferring it to a database system is usually regarded as time consuming. Also, it is common for the physicians in different hospitals to have slightly different investigation methods. This results in different structures for the data coming from different sources, making it impossible to combine it in most cases. Moreover, since human life is at stake, accurate diagnosis is crucial. Establishing the baseline accuracy for a given dataset is therefore very important as well. Domain knowledge or ethical and social issues are also of great significance.

An essential particularity of medical problems is the concept of *cost*, which is addressed by cost-sensitive classification. This idea will be developed further in Section 2.

This paper tries to tackle several issues involved in medical data mining. Section 2 introduces ProICET, a cost sensitive approach to the medical diagnosis process. The system introduced in Section 3 tries to improve the accuracy of symbolic classifiers, while keeping the diagnosis process transparent to the physician. Section 4 presents a system for classifier combination, based on the Dempster-Shafer Theory of evidence combination. The system can be employed to establish the baseline accuracy of a dataset, and to combine medical data coming from different sources, having different structure.

2 ProICET – A Cost-Sensitive System for Medical Diagnosis

2.1 Theoretical Aspects

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 of 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.

The concept of cost is addressed in a separate area of machine learning, known as cost-sensitive learning. There are two main categories of cost-sensitive techniques: algorithms that are sensitive to misclassification costs (stratification, MetaCost, AdaCost), and algorithms that consider test costs (EG2, CS-ID3, IDX). Significantly less work has been done for aggregating several cost components. The most prominent approach in the literature is ICET, first developed by Peter D. Turney [11].

The ICET algorithm takes on a hybrid approach to considering costs in the learning process: it combines a greedy search heuristic (Eg2) with evolutionary means (genetic algorithms). The algorithm can be viewed as working at two levels:

- On the bottom level, Eg2 performs a greedy search in the space of decision trees
- On the top level, the evolutionary component performs a genetic search through a space of biases; the biases are used to control Eg2's preference for certain types of decision trees.

Some enhancements have been considered in the genetic component. The most important 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 (this reduces the search variability).

2.2 Evaluation Methodology

The experiments performed on ProICET have been carried out in two phases. In a first phase, we wanted to verify whether this particular implementation of the ICET algorithm performed better in real cases than other similar algorithms; we also wanted to compare the results with the initial implementation of the algorithm. In order to do so, we planned three series of tests: the first one tried to solve a problem that Turney himself discovered, related to the asymmetry in the costs for rare examples; a second set of tests provided a more comprehensive analysis of the misclassification cost component; a third concern was to study the behaviour of ProICET in real world conditions (medical diagnosis problems with real costs).

The datasets used for these tests have been obtained from the machine learning data repository website. All of them are from the medical field. Most of the datasets were also used in the original work on ICET [11].

In the second phase we studied the behaviour of the ProICET system on real prostate cancer data, and compared its performances with those of the algorithms included in the first phase.

For both phases we used the following genetic parameters setup: 1000 evaluation steps, Gray coding for individual representation, 14 genes in a chromosome, 50 individuals in the population, roulette wheel as the parent selection method, multiple point crossover, with 4 random points, single point mutation, with one random point, and 0.2 mutation rate. Due to the strong heuristic component, every result has been averaged over 10 runs. Each run used a 70-30% split.

In order to study the cost asymmetry and the misclassification cost component, we used a two class problem and the following cost matrix:

$$C = 100 \cdot \begin{pmatrix} 0 & p \\ 1-p & 0 \end{pmatrix} \qquad (1)$$

where p was varied between 0 and 1, with 0.05 increments.

For both phases we have employed the following algorithms in order to provide a comparative study: J4.8 (revision 8 of the C4.5 implementation found in Weka), AdaBoost.M1, MetaCost, and Eg2.

In phase two we did not have any cost settings for the prostate cancer dataset. Consequently, we used two values for the test costs: 0 and 0.1, and the cost matrices presented in Table 1 (M stands for matrix). In setting the misclassification costs we experimented on several different values for the unbalance in the error costs, while keeping the same magnitude order.

Table 1 – Cost matrices for the prostate cancer dataset

M 1	lo	med	hi	M 2	lo	med	hi
lo	0.0	0.5	1.0	lo	0.0	0.5	1.0
med	1.5	0.0	0.7	med	3.0	0.0	0.7
hi	5.0	3.0	0.0	hi	10.0	6.0	0.0
M 4	lo	med	hi	M 3	lo	med	hi
lo	0.0	0.5	1.0	lo	0.0	0.5	1.0
med	3.0	0.0	0.5	med	0.75	0.0	0.7
hi	5.0	3.0	0.0	hi	2.5	1.5	0.0

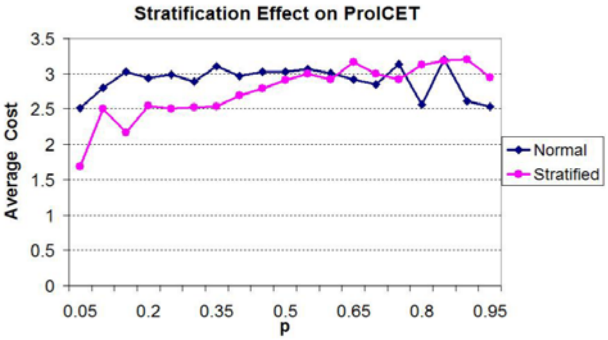

Figure 1 – Stratification effect on the average cost of ProICET for the Wisconsin dataset

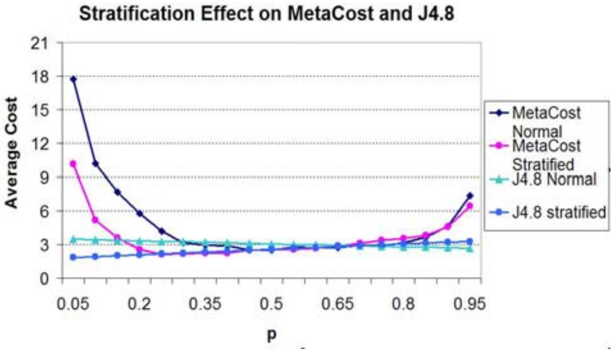

Figure 2 – Stratification effect on the average cost of MetaCost and J4.8 for the Wisconsin dataset

2.3 Results

As mentioned above, a first concern in the initial evaluation stage was to check if stratification could improve the cost characteristic of classifiers for the rare, expensive cases. The method consists in altering the distribution of examples for each class, such as to include proportionally more examples of the classes having high misclassification costs.

The effects of stratification on ProICET for the Wisconsin dataset are presented in Figure 1. Wisconsin has been chosen for this experiment because it is one of the largest, two-class, medical dataset. We can observe a small decrease in misclassification costs for the stratified case throughout the parameter space. This reduction becomes significant at the margins, and especially in the left margin, where the rare expensive cases are.

Figure 2 illustrates the effects of stratification on the costs obtained by J4.8 and MetaCost on the Wisconsin dataset. Here also we observe a significant reduction in the left part of the chart for the stratified case.

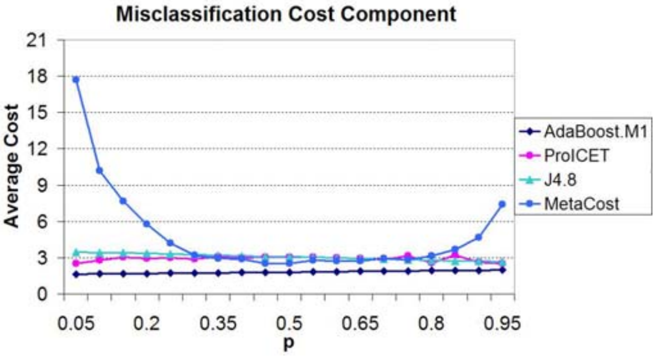

Figure 3 – Misclassification cost analysis on the Wisconsin dataset

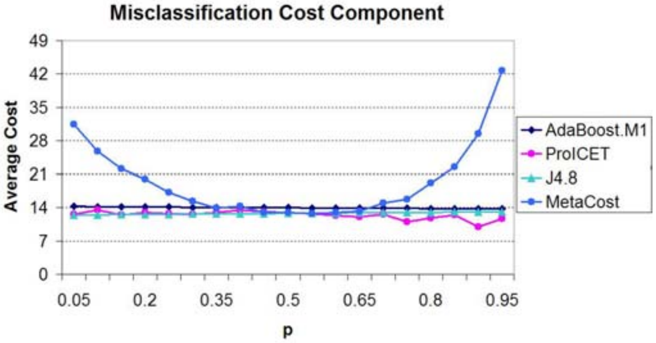

Figure 4 – Misclassification cost analysis on the Pima dataset

We conclude that stratification could be used for improving the cost characteristic of some classifiers. Further testing is required before formulating more general results.

The second concern of the first evaluation phase was related to studying the misclassification cost component of several algorithms. For this we used two-class medical problems (Wisconsin and Pima datasets).

As illustrated by Figure 3, MetaCost yields the poorest results on the Wisconsin dataset. ProICET performs slightly better than J4.8, while the smallest costs are obtained for AdaBoost.M1, using J4.8 as base classifier. The fact that AdaBoost obtains lower costs than ProICET can be explained by the different approaches taken when searching for a solution. If ProICET uses heuristic search, AdaBoost implements a procedure that is guaranteed to converge to minimum training error, while the ensemble voting reduces the risk of overfitting.

This behaviour changes on the Pima dataset (Figure 4), where ProICET yields slightly lower costs than AdaBoost.M1. The performance of MetaCost is poor in this case as well.

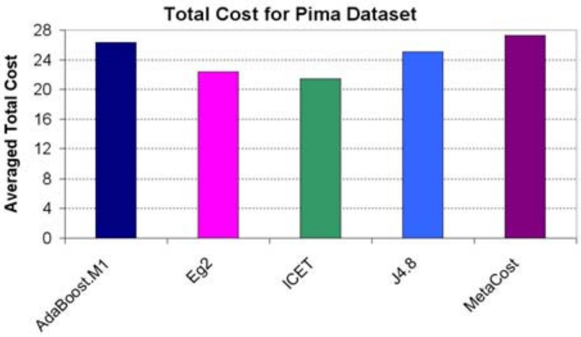

Figure 5 – Total cost for the Pima dataset

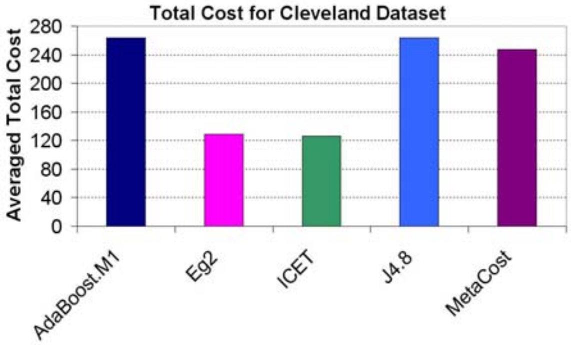

Figure 6 – Total cost for the Cleveland dataset

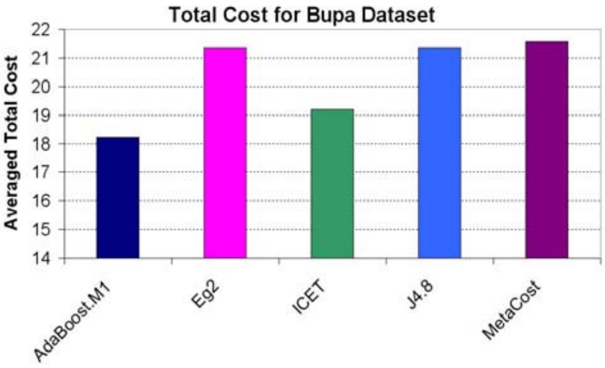

Figure 7 – Total cost for the Bupa dataset

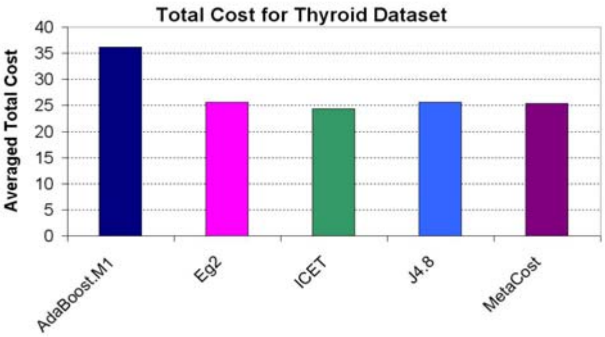

Figure 8 – Total cost for the Bupa dataset

In real world situations we expect ProICET to perform better than the other algorithms, since it is the only approach that takes into account both test and misclassification costs. Indeed, figures 5-8 show that ProICET is the best at reducing the total cost. An interesting result is obtained on the heart disease Cleveland dataset, where the relative cost reduction is more than 50%. This is a significant improvement to the initial implementation of ICET, and is probably owed to the alterations made to the genetic algorithm, which increase population variability and extend the ICET heuristic search.

The second evaluation phase was concerned with evaluating ProICET on a real prostate cancer dataset. The problem of interest was to predict the class of postoperative PSA (low, medium, high) from preoperative (preoperative PSA, Gleason score, prostate volume, International Index of Erectile Function, quality of life, TNM) and operative parameters (surgery type, operative technique, nerve sparing, bleeding, operative time, a.s.o.).

The focus here was on both the cost values and the accuracy rates. We expected that ProICET yield low costs while keeping high accuracy rates.

The accuracy rates, averaged over the eight different cost settings (two different values for the test costs and four cost matrices), can be observed in Figure 9. ProICET yields the highest accuracy rates, together with Eg2 and J4.8. The fact that the accuracy is not as high as expected (around 84%), could be rooted in the characteristics of the dataset (relatively small number of instances, some missing values). A method for increasing the volume of the data and improving the quality is currently investigated. It is based on applying ensemble learning methods to artificial neural networks (more in Section 3).

Table 2 illustrates the total costs obtained by the algorithms over the various cost settings considered. When both types of costs are involved, ProICET yields lowest total cost (bolded in the table).

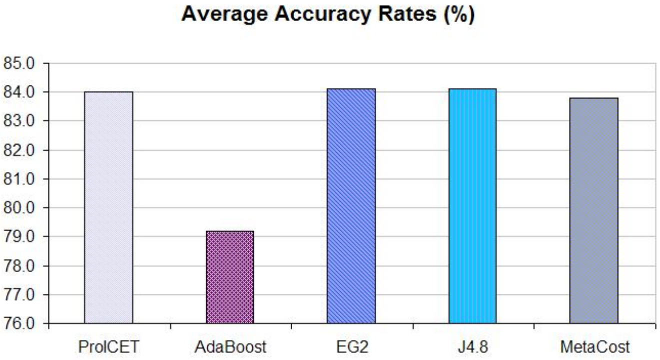

Figure 9 – Average accuracy rates for the prostate cancer dataset

Table 2 – Total cost for the prostate cancer dataset

Average Total Cost					
	Pro ICET	Ada Boost	Eg2	J48	Meta Cost
Test 0 Matrix 1	0.28	0.284	0.269	0.269	0.293
Test 0.1 Matrix 1	**0.414**	0.734	0.430	0.430	0.448
Test 0 Matrix 2	0.561	0.52	0.52	0.52	0.65
Test 0.1 Matrix 2	**0.678**	0.97	0.682	0.682	0.812
Test 0 Matrix 3	0.146	0.166	0.142	0.142	0.145
Test 0.1 Matrix 3	**0.252**	0.616	0.305	0.305	0.310
Test 0 Matrix 4	0.213	0.44	0.44	0.44	0.502
Test 0.1 Matrix 4	**0.575**	0.89	0.603	0.603	0.647

Also, ProICET yields the lowest overall total costs, with an impressive 30% relative cost reduction when compared to AdaBoost.M1. This proves once again that ProICET is the best at reducing the costs involved in medical problems.

A medical result obtained in this phase is related to the ranking of the predictor attributes. By analyzing the output trees of several methods involved in the evaluation, we found the following order:

1. Prostate Volume
2. Operation Technique
3. Bleeding
4. Gleason Score
5. IIEF (International Index of Erectile Function)
6. Preoperative PSA

The newly discovered importance of the prostate volume is yet to obtain medical recognition, but its significance has been suspected by some physicians.

3 PANE Method (Preceded by Artificial Neural Network)

3.1 Theoretical Aspects

As mentioned before, a high accuracy rate is crucial when employing data mining methods to support medical diagnosis and prognosis. Moreover, the decision process must be transparent enough such as to allow the physician to understand the rationale behind the verdict indicated by the system.

Many machine learning approaches fail to meet both the transparency and the performance requests. While the symbolic techniques, such as decision trees and rule learners, represent knowledge in a structured way, easily comprehensible by the user, they don't have the improved performance of the connectionist methods (artificial neural networks). The latter encode the information inside their architecture, making them less transparent and harder to comprehend. Besides the transparency issues caused by their connectionist nature, artificial neural networks are also unstable, that is small changes in the training data may result in very different models, thus affecting the performance on unseen data. [14]

The PANE method (Preceded by Artificial Neural Network Ensemble) tries to address both the transparency and the stability problems, by using a neural network ensemble as a preprocessing step for a symbolic classifier. Ensemble learning methods are known to improve the generalization abilities of simple classifiers, therefore increasing their stability. The transparency is ensured by the use of a symbolic classifier to obtain the output model. This method tries to satisfy both the comprehensibility and the robustness needs involved in medical diagnosis and prognosis.

We have implemented a system based on the PANE method, and evaluated it on several benchmark datasets. The system flow can be split into the following steps:

1. Train the neural network ensemble using the available training data.
2. Use the trained ensemble to classify the available data.
3. [*Optional*] Generate some random data and use the trained ensemble to classify it.
4. Run a symbolic learning algorithm on the data generated in steps 2 and 3.

In step 1, the artificial neural network ensemble can be obtained in several ways: bagging with un-weighted majority vote, bagging with weighted majority vote, and boosting. For the symbolic classifier we have used three algorithms: C4.5 (revision 8 of the algorithm, found in the Weka framework – J4.8), PART (a rule learner) [2] and AdaBoost.M1. Although AdaBoost.M1 is not a symbolic classifier, we have included it in our work to better study the impact of the neural network ensemble on the performance of simple classifiers. Also, there exist methods for extracting rules from ensembles of trees, therefore making it practical in real situations also.

3.2 Evaluation Methodology

The evaluations performed tried to confirm that the idea of preceding symbolic classifiers with artificial neural network ensembles is effective and can improve the performance of the symbolic classifiers, while keeping their transparency. A second concern is associated with the idea formulated in [15], regarding the positive impact that random data can have on the learning capabilities of the symbolic classifier (the optional step 3 in the system). The procedure presented there is very simple: data is generated randomly and is fed to the neural network ensemble, which predicts class labels for each feature vector. The newly obtained data is combined with existing, relabeled training set and provided to the symbolic learner for training. In [15] it was concluded that this was a good technique for increasing the accuracy of the symbolic classifier.

We have used five benchmark datasets, obtained from the UCI Machine Learning Repository: Bupa liver disorder, Cars, Cleveland heart disease, Pima Indian diabetes and Wisconsin breast cancer. Four of the datasets are from the medical domain, and one is from the car manufacturer industry.

Results have been averaged over 100 runs. For each run, 80% of the available instances have been used for training, and the rest of 20% have been kept for evaluation. The parameters of the artificial neural networks have been set to the following values: 1 hidden layer, (number of attributes + number of classes)/2 for the number of units in the hidden layer, 0.6 for the learning rate and 0.2-0.3 for the momentum. We have also used a validation set in the training of the neural networks, composed of instances which were not present in the bootstrap sample used for training.

3.3 Results

The experimental results using bagging with un-weighted majority vote are shown in Table 3. We have used the following abbreviations: NE for the algorithm enhanced by the neural network ensemble and NNE for neural network ensemble alone. Significant error reductions can be observed for all the classifiers, on all the datasets. On average, PART achieves the highest relative error reduction (12%), with an impressive 32% on the Cars dataset.

The results for bagging using weighted majority vote (Table 4) are quite similar to those obtained with un-weighted majority vote, except for the Cars dataset, where the improvement is not as significant as before.

Table 3 – Error rates obtained with bagging using un-weighted majority vote

Algorithm	*Bupa (%)*	*Cleveland (%)*	*Pima (%)*	*Wisconsin (%)*	*Cars (%)*
C4.5	39.38	47.35	26.58	5.81	8.85
C4.5 NE	36.70	45.56	24.88	5.11	7.20
PART	38.45	48.83	27.07	4.32	4.87
PART NE	36.16	45.11	25.10	4.00	3.31
AdaBoost	39.44	47.46	27.22	3.48	4.87
AdaBoostNE	36.02	44.42	23.94	3.35	4.86
NNE	35.25	42.02	23.73	3.31	0.66

Table 4 – Error rates obtained with bagging using weighted majority vote

Algorithm	*Bupa (%)*	*Cleveland (%)*	*Pima (%)*	*Wisconsin (%)*	*Cars (%)*
C4.5	39.69	48.13	26.67	5.20	8.85
C4.5 NE	36.55	45.39	25.05	4.72	8.77
PART	38.36	47.61	27.37	5.94	4.99
PART NE	36.20	45.93	25.32	5.91	4.91
AdaBoost	38.55	49.56	25.65	4.66	5.91
AdaBoostNE	34.30	49.50	24.32	4.55	5.59
NNE	34.85	42.18	24.07	3.58	0.73

Table 5 – Error rates obtained with boosting

Algorithm	*Bupa (%)*	*Cleveland (%)*	*Pima (%)*	*Wisconsin (%)*	*Cars (%)*
C4.5	40.36	47.25	26.68	5.45	8.82
C4.5 NE	38.30	44.70	25.69	4.66	20.29
PART	38.27	48.23	27.20	4.60	5.04
PART NE	36.38	43.77	25.30	4.07	16.13
AdaBoost	38.26	48.36	25.49	5.05	5.24
AdaBoostNE	35.88	45.89	24.96	3.76	29.34
NNE	35.16	42.60	24.62	3.77	2.35

Table 5 shows the results obtained using boosting as the method for constructing the neural network ensemble. The overall relative reduction in the error rate is about 8%. An important remark should be made here about the poor performance of the PANE method on the Cars dataset.

This happens because the boosting process stops when the error goes below a given threshold (1%). On the Cars dataset, the neural network ensemble reaches this threshold in 1-2 iterations. Thus, the ensemble will only have 1-2 members, its generalization ability being therefore affected. This proves once again the idea formulated in the "No Free Lunch" theorem, which states that there is no algorithm which performs better than all the other algorithms, on all the existing problems. This issue will be further addressed in Section 4. An interesting result would be to find a rule for setting the error threshold automatically, such as to maintain the ensemble diversity, but still avoid overfitting.

The impact that additional random data has on the performance of C4.5 used with bagging (un-weighted majority vote) can be observed in figures 10-13. The values on the x axis represent the size of the randomly generated data, being a multiple of the initial set size.

A first remark is that additional training data can reduce the error rate of the symbolic classifier. The value of the ratio of additional data for which the best performance is obtained depends on the dataset. We observed that increasing the data with an amount equal to 4 or 5 times the initial dataset size usually leads to a significant relative error reduction (up to 70% for the Cars dataset). However, increasing this ratio further may not lead to improved performance.

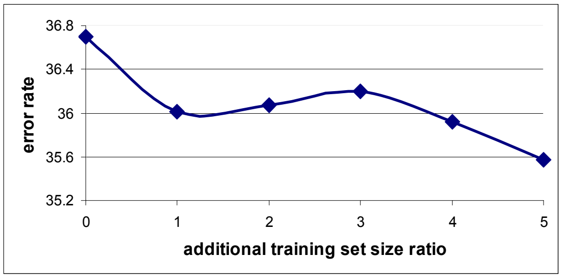

Figure 10 – Error variation for different sizes of the additional data on Bupa

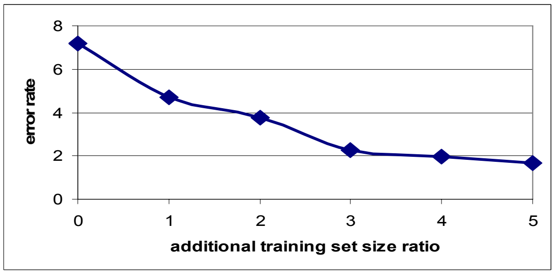

Figure 11 – Error variation for different sizes of the additional data on Cars

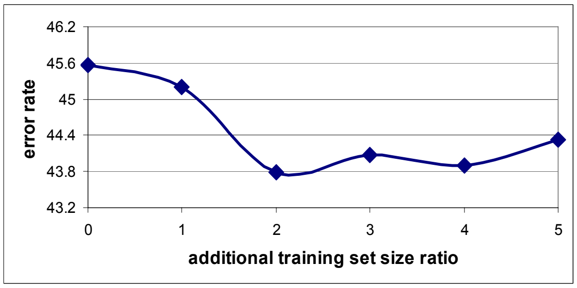

Figure 12 - Error variation for different sizes of the additional data on Cleveland

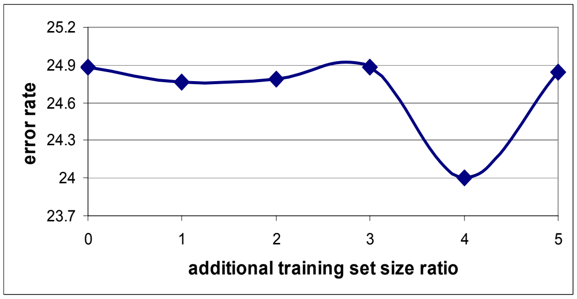

Figure 13 – Error variation for different sizes of the additional data on Pima

Since the neural network ensemble generally yields lowest error rates, and the addition of random data has proved to be beneficial, neural network ensembles and the method of generating data could be used to increase the volume of the available data and improve its quality.

4 Classifier Combination through the Dempster-Shafer Theory

Besides the accuracy and data-related issues, specific to the medical domain, this section addresses some issues of machine learning algorithms which use a single hypothesis. They are known to suffer from several drawbacks: the statistical problem, or high variance, which occurs when the hypothesis space is too large for the available training data; the computational problem, or computational variance, which occurs when the learning algorithm cannot guarantee to find the best hypothesis within the computational space; and the representation problem, or high bias, which occurs when none of the hypotheses in the search space is a good enough approximation of the truth.

While ensemble methods can reduce both the bias and the variance of learning algorithms, they do not solve the problem of failing to choose a classifier that will perform best for a given dataset. In addition, there is also the problem of establishing a lower bound to the accuracy on a certain problem. This is what we try to achieve with the classifier fusion method proposed by the mathematical theory of Dempster and Shafer.

The Dempster-Shafer Theory is a theory of evidence, based on belief functions and plausible reasoning. Its main feature is that it combines several pieces of information in order to compute the probability of an event. Moreover, it allows for directly representing the uncertainty of system responses: the imprecise input can be modeled by a set or an interval, and the output is a set or an interval. Initial efforts for developing the theory were made by A. Dempster (1967), but the theory was completed by the seminal work performed by G. Shafer (1976) [10].

The system is intended to provide a reference accuracy value when choosing a classifier for any specific dataset. Due to the advantages provided by the fusion technique, the risk is minimized and the accuracy obtained by applying the combined classifier on any data is surely among the highest possible. The classifiers used in the combination are the Bayesian Classifier, k-Nearest Neighbour and Decision Tree learner.

There are three main steps in designing the system, namely belief extraction from the three classifiers, uncertainty computation, and belief combination:

1. *Extracting beliefs from the three classifiers* – takes into account the nature of each classifier. The following approaches have been used:
    - For the Bayesian classifier, the posterior probability function is used for belief evaluation.
    - For k-Nearest Neighbour, a distance function is used to evaluate basic beliefs.
    - For the decision tree, the confidence is the measure for evaluating beliefs.
2. *Computing the uncertainty for the classifiers* – evaluate the distance between the belief value and the value $1/K$ (where K is the number of classes). Uncertainty is then computed as the normalized sum of the square of these distances; the closer the sum is to zero, the higher the uncertainty.
3. *Combining the evidence* - combine the belief and the uncertainty obtained in the previous steps, such as to arrive at the final decision.

In our implementation, the Bayesian classifier is first combined with the kNN classifier, and then the resulting classifier is combined with the decision tree learner. The combination takes advantage of the fact that one classifier may be more accurate in handling records corresponding to a certain class than the other.

4.1 Evaluation Methodology

The evaluation focused on obtaining validation that the combination of the classifiers using the Dempster-Shafer theory is robust and stable and can be used to asses baseline accuracies for any dataset.

For this, we have used four benchmark datasets, obtained from the UCI Machine Learning Repository: Cars, Cleveland heart disease, Pima Indian diabetes and Wisconsin breast cancer.

The testing methodology assumed averaging the accuracy over 100 runs; for this, each dataset was used to generate 100 random pairs of training/testing datasets (using 80/20 percentage split). We have performed tests for each of the three classifiers separately, on all datasets; then, the combined classifier was tested on the datasets and the results were compared.

For a better validation of the system, comparisons with ensemble learning methods have been carried out (bagging and boosting in combination with the three classifiers involved in the evaluation).

4.2 Results

As Table 6 shows, the individual classifiers are not stable with respect to the datasets. While Naïve Bayes seems to classify the instances in the *Wisconsin* dataset most accurately, it performs poorly on the *Cars* dataset, where it produces the lowest accuracy among the three classifiers.

Table 6: Individual classifiers accuracies

Dataset	*Bayes*	*kNN*	*J4.8*
Cars	85.43%	92.30%	91.50%
Cleveland	55.73%	56.91%	52.60%
Pima	75.44%	73.38%	73.88%
Wisconsin	96.24%	95.35%	94.41%

Table 7: Comparison between the accuracy of the combined classifier and the average accuracy of the three classifiers

Dataset	*Average*	*DST combined classifier*
Cars	89.74%	91.55%
Cleveland	55.08%	55.81%
Pima	74.23%	74.85%
Wisconsin	95.33%	96.16%

Similar remarks can be done for the other two classifiers. Another remark is related to the fact that the differences in accuracy between the three classifiers are high, especially in the case of *Cars* datasets, which proves the high value of the risk involved in choosing a certain classifier for a certain dataset.

From Table 7 we can observe that the combined classifier is more accurate than the average of the three classifiers. Even if a classifier performs better on a certain dataset than the combined classifier, the same classifier will perform poorly on other datasets. For example, in the case of the *Wisconsin* dataset, the Bayesian classifier yields highest accuracy, while on the *Cars* dataset it achieves the poorest performance among the three classifiers. These results are in strong connection with the “No Free Lunch” theorem.

Tables 8-10 present the results obtained by bagging and boosting (tables 8 and 9) and the comparison with the combined classifier (Table 10). It can be observed that even though the ensemble learning methods improve the accuracy, the problem of the differences between the classifiers’ predictions on a dataset still persists, especially in the case of bagging. The same observation can be made about the problem of a classifier being the best predictor on one dataset and the worst on another.

Table 8: Accuracy rates for bagging

Dataset	Bagging +Bayes	Bagging +kNN	Bagging +J4.8
Cars	85.14%	93.10%	92.71%
Cleveland	55.91%	58.01%	54.26%
Pima	75.38%	73.48%	75.11%
Wisconsin	97.40%	95.61%	95.33%

Table 9: Accuracy rates for boosting

Dataset	Boosting +Bayes	Boosting +kNN	Boosting +J4.8
Cars	90.35%	92.30%	95.21%
Cleveland	55.73%	53.55%	53.18%
Pima	75.61%	73.33%	72.31%
Wisconsin	95.68%	95.23%	96.24%

Table 10: Accuracies for the combined classifier and the ensemble learning methods

Dataset	Bagging	Boosting	DST Combined classifier
Cars	90.32%	92.62%	91.55%
Cleveland	56.06%	54.15%	55.81%
Pima	74.66%	73.75%	74.85%
Wisconsin	96.11%	95.72%	96.16%

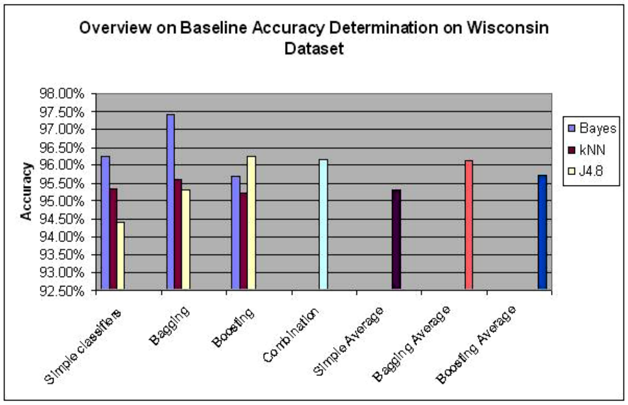

Figure 14 – Accuracy baseline determination for the Wisconsin dataset

The combined classifier outperforms bagging on three datasets and boosting on other three datasets, being slightly less accurate than one of the three classifiers on one dataset. This proves yet again the risk minimization displayed by the combined classifier.

The chart in Figure 14 illustrates the idea of using the combined classifier to set the baseline accuracy for a certain dataset. As the chart shows, only three classifiers outperform the combined classifier: Naïve Bayes, bagging with the Bayesian classifier and boosting with the decision tree classifier. Also, the improvement is not always significant enough. Moreover, there is no guarantee that the same classifiers will perform equally well on another dataset.

The combination method proposed by the Dempster-Shafer theory has another advantage: it can combine data coming from different sources, with different structures. This cannot be achieved through bagging or boosting. As mentioned before, being able to combine data coming from separate sources is particularly useful in the medical domain, where hospitals may have slightly different investigation methods.

5 Conclusions

Among the diverse domains in which data mining methods have been applied to support the decision process, the medical field poses some the most challenging particularities. Some are data related; some refer to domain knowledge and ethical and social issues. A high accuracy rate is essential for any diagnosis. The concept of cost is also of great importance.

This paper tries to address some of these issues, by proposing several systems. A hybrid, cost-sensitive approach has been adopted for ProICET. The evaluations performed confirmed ProICET achieves lowest total costs, while keeping high accuracy rates. This could provide valuable support to reducing the economical and time-related costs involved in the diagnosis process, and help avoid "dangerous" misclassifications.

A second system implements a PANE method to achieve both high accuracy rates and improved comprehensibility. The low error rates achieved by the neural network ensemble, and the promising results obtained by the method used to artificially increase the volume of the data have lead to the initiative of using these methods to improve the data quantity and quality. Currently we are investigating a procedure for enhancing medical data, following this idea.

A method for combining classifiers, based on the data fusion principles, is employed to build a third system. This idea has proved to be beneficial for several reasons. First of all, the system can be used to provide the baseline accuracy for any dataset, thus reducing the risk of choosing an inappropriate classifier. Only methods which yield higher accuracy rates than the combined classifier are considered proper for the dataset at hand. Secondly, the theoretical basis of the method allows for data coming from different sources to be combined. This is particularly important in the medical domain, since it could unify data coming from different hospitals, having slightly different structures. Evaluations are being conducted to validate this idea as well.

Our present work focuses on combining data coming from different sources, increasing the volume of the data, and predicting missing values for the attributes. The development of a system dealing with these issues is currently under research.

A future goal is to integrate the systems presented here, as well as future components, into a cohesive framework, to support medical data collection, processing, interpretation and the decisions involved in the diagnosis process.

References

[1] L. Breiman, "Bagging predictors", *Machine Learning*, 24, 1996, pp. 123-140

[2] E. Frank and I. Witten. "Generating Accurate Rule Sets without Global Optimization", *Machine Learning: Proceedings of the Fifteenth International Conference*, Morgan Kaufmann Publishers, San Francisco, 1998.

[3] Y. Freund, R. E. Schapire, "Experiments with a New Boosting Algorithm", *International Conference on Machine Learning*, 1996.

[4] L. K. Hansen and P. Salamon. "Neural network ensembles", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.12, no.10, pp.993-1001,1990.

[5] G Mahajani, Y Aslandogan, "Evidence Combination in Medical Data Mining", University of Texas, Arlington, USA, 2003.

[6] T. Moldovan, C. Vidrighin, I. Giurgiu and R. Potolea, "Evidence Combination for Baseline Accuracy Determination". Accepted at ICCP 2007, Cluj-Napoca, Romania

[7] A. Onaci, C. Vidrighin, M Cuibus and R. Potolea, "Enhancing Classifiers through Neural Network Ensembles". Accepted at ICCP 2007, Cluj-Napoca, Romania.

[8] R. Polikar, D. Parikh, S. Mandayam, "Multiple Classifier Systems for Multisensor Data Fusion", *IEEE Sensors Applications Symposium*, Houston, Texas, USA, 2006.

[9] R. Potolea, C. Vidrighin, C. Savin. ProICET – A Cost-Sensitive System for the Medical Domain. Accepted at ICNC'07-FSKD'07, Haikou, China, August 2007

[10] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976.

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

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

[13] C. Vidrighin, C. Savin and R. Potolea, "A Hybrid Algorithm for Medical Diagnosis". Accepted at the IEEE Region 8 Eurocon 2007 Conference, September 2007.

[14] R. Wall and P. Cunningham, "Exploring the potential for rule extraction from ensembles of neural networks", *11th Irish Conference on Artificial Intelligence & Cognitive Science*, 2000.

[15] Z.H. Zhou and Y. Jiang, "Medical Diagnosis with C4.5 Rule Preceded by Artificial Neural Network Ensemble', *IEEE Transactions on Information Technology in Biomedicine,* 2002.