Controlling the Cost of Prediction in using a Cascade of Reject Classifiers for Personalized Medicine

Blaise Hanczar[1] and Avner Bar-Hen[2]
[1]IBISC, IBGBI University of Evry, 23 bd de France, 91034 Evry, France
[2]MAP5, University Paris Descartes, 45 rue des saint-peres 75006 Paris, France

Keywords: Supervised Learning, Reject Option, Cascade Classifier, Genomics Data, Personalized Medecine.

Abstract: The supervised learning in bioinformatics is a major tool to diagnose a disease, to identify the best therapeutic strategy or to establish a prognostic. The main objective in classifier construction is to maximize the accuracy in order to obtain a reliable prediction system. However, a second objective is to minimize the cost of the use of the classifier on new patients. Despite the control of the classification cost is high important in the medical domain, it has been very little studied. We point out that some patients are easy to predict, only a small subset of medical variables are needed to obtain a reliable prediction. The prediction of these patients can be cheaper than the others patient. Based on this idea, we propose a cascade approach that decreases the classification cost of the basic classifiers without dropping their accuracy. Our cascade system is a sequence of classifiers with rejects option of increasing cost. At each stage, a classifier receives all patients rejected by the last classifier, makes a prediction of the patient and rejects to the next classifier the patients with low confidence prediction. The performances of our methods are evaluated on four real medical problems.

1 INTRODUCTION

The personalized medicine is an ongoing revolution in medicine; its objective is to maximize the wellness for each individual rather than simply to treat disease. According to Hood and Friend (Hood and Friend, 2011), this revolution is based on several points. The first one is to consider that medicine is an information science. The second point is the emergence of technologies that will let us explore new dimensions of patient data space, like the ”omics” technologies. The last point is the development of powerful new mathematical and computational methods, specially in machine learning, that will let us analysis the marge amount of data associated with each individual. Today physicians have access to a large amount data for each patient from different sources: clinical, environmental, psychological, biologic or omic. The use of automated methods is indispensable to analyse and extract relevant information from these data. An important way of research is the development of prediction systems whose objectives generally are to diagnose a disease, to identify the best therapeutic strategy or to establish a prognostic for a patient. These systems, called classifiers, are constructed from supervised learning methods, the most popular are the discriminant analysis (Dudoit et al., 2002), the support vector machine (Furey et al., 2000), the random forest (Diaz-Uriarte and Alvarez de Andres, 2006), the neural networks (Khan et al., 2001) or the ensemble methods (Yang Pengyi; Hwa Yang Yee; Bing B. Zhou;, 2010).

The primordial objective of the prediction systems is to maximize their accuracy in order to obtain reliable predictions. However, a second objective, generally ignored in research studies, is to minimize the cost of the prediction. In a classifier, a patient is represented by a set of variables. These variables come from different medical exams and each of these exams has a cost. The use of a classifier requires the values of all variables of the patient; the cost of the prediction is the sum of the costs of all exams used by the classifier. Note that the cost does not necessary represent money, it may also represent time, secondary effects of treatment or any other non-infinite resource. In practice, a good prediction system has to both maximize its accuracy and minimize its cost.

In this paper we propose a new method that reduce the prediction cost without increasing the error rate. In prediction problems, it is worth to note that some patients are easier to predict than others and do not need all medical exams. For these patients, a reliable prediction can be done with a small sub-

Hanczar, B. and Bar-Hen, A.
Controlling the Cost of Prediction in using a Cascade of Reject Classifiers for Personalized Medicine.
DOI: 10.5220/0005685500420050
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2016) - Volume 3: BIOINFORMATICS*, pages 42-50
ISBN: 978-989-758-170-0

set of variables and can be therefore less expensive. Based on this observation, we propose a new supervised classification approach using a cascade of classifiers with reject option. This cascade is a sequential set of classifiers with reject option of increasing cost. The patient data are submitted to the first classifier that makes a prediction. If this prediction is judged not reliable, the patient is rejected to the next classifier of the cascade needing additional variables. The process is repeated until a reliable prediction has been done. This approach allows reducing the cost of the basic classifier using all variables. In this approach, there is a trade-off between the accuracy and the cost of the predictions. The two main scientific keys of our method are the computation of the rejection area of all classifiers of the cascade. The second is to find the optimal order of the variables that form the structure of the cascade. The section two gives the state of the art of cost minimization methods and cascade classification. In section three, we provide the formulation of the classification with rejection option and the cascade. The two algorithms for the computation of the rejection areas and the order of the variables are given in detail. The section four presents the results on four real datasets and analysis the performance of our method.

2 RELATED WORK

The reduction prediction cost problem is close to the active feature acquisition problem in the cost sensitive learning (Saar-Tsechansky et al., 2009). The objective is sequentially decided if we want to acquire the next feature in order to increase the accuracy of the classifier. Markov decision process is one of the usual approach used in this context, for example, Kapoor (Kapoor and Horvitz, 2009) propose a new class of policies inspired from active learning. Tan propose an attribute value acquisition algorithm driven by the expected cost saving of acquisition only for support vector machine (Tan and yen Kan, 2010). Nan developed a variant of random forest dealing with the cost of the variables (Nan et al., 2015).

One simple structure for incorporating the cost into learning is through a cascade of classifiers. This approach has been popularized by Viola and Jones (Viola and Jones, 2004) with their detection cascade used in image analysis for object detection. Cheap variables are used to discard examples belonging to the negative class. This type of method is focused on unbalanced data with very few positive examples and a large number of negative examples. Note that the main objective of this Viola's cascade is to increase the accuracy of the prediction, they do not deal with the prediction cost. In the context of information retrieval, Wang adapted the cascades to ranking and incorporated variables costs but retained the underlying greedy paradigm (Wang et al., 2011). Raykar et al. (Raykar et al., 2010) explore the idea of a cascade of reject classifier. Their version is a soft cascade where each stage accepts or rejects examples according to a probability distribution induced by the previous stage. Each stage of the cascade is limited to linear classifiers, but they are learned jointly and take in account of the cost of the variables. Trapeznikov and Saligrama (Trapeznikov and Saligrama, 2013) propose a multi-stage multi-class system where the reject decision at each stage is posed as a supervised binary classification problem. They derive bound for VC dimension to quantify the generalization error.

A common limitation of all these methods is that the order of the variables is supposed to be known. The structure of the cascade is therefore fixed. Our method overcomes these limitations in using a heuristic to compute an order of the variables.

3 CASCADE OF REJECT CLASSIFIERS FORMULATION

3.1 Formulation of the Problem

We consider a classification problem with two classes (positive "1" and negative "0") with D variables $\{v_1,...,v_D\}$. Let's a training set of N examples $T = \{(x_1,y_1),...,(x_N,y_N)\}$ where $x_i \in \mathbb{R}^D$ is the variable vector and $y_i \in \{0,1\}$ is the label. We denote c_i the cost for acquiring the i-th variable of an example. Let's $\Psi : \mathbb{R}^D \to \{0,1\}$ the basic classifier constructed from a usual supervised learning procedure and making predictions in using all variables. Our objective is to construct a cascade that obains better performances than the basic classifier.

In this context, the performance of a classifier is measured by two values: its error rate i.e. the probability that the prediction does not correspond to the true label, noted $E = p(\Psi(x) \neq y)$ and its cost that is the total acquisition cost of all variables required by the classifier, noted $C = \sum_{i=1}^{d} c_i$. These values are combined into a new value called loss, that represent the total performance of the classifier and is defined by:

$$L = C + \Lambda E \qquad (1)$$

Λ is a parameter that represented the penalty of a misclassification. In our cascade, this parameter controls the trade-off between the cost and the error rate. For

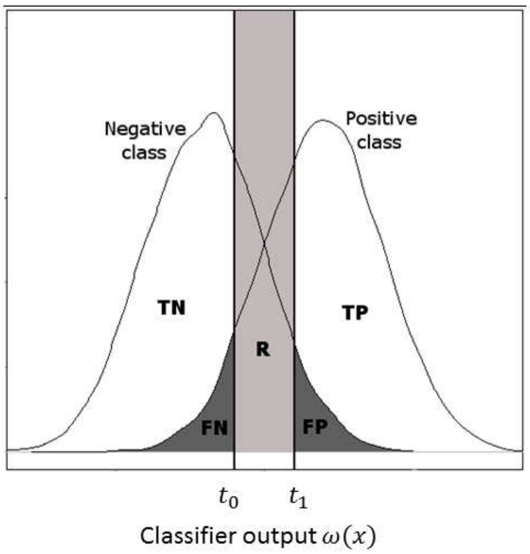


Figure 1: Distribution of the classes on the classifier output. TP, TN, FP, FN and R represent respectively the true positive, true negative, false positive, false negative and rejection area.

the basic classifier, the cost C is constant since we always have to pay all variables. The objective is to construct a cascade with a loss lower than the loss of the basic classifier.

3.2 Classifier with Rejection Option

The base element of our cascade system is the classifier with reject option. This type of classifier can reject examples if it is not enough confident in the predictions. No class is assigned to rejected examples. Let's Ψ a classifier whose the output $\omega(x)$ is a continuous value. In fixing a threshold t on this output, we define a classic classifier that assigns one of the two classes to each example. In fixing two thresholds $\{t_0, t_1\}$, we define a classifier that rejects some examples and assigns one of the two classes to the non-rejected examples.

$$\Psi(x) = \begin{cases} 0 & \text{if} \quad \omega(x) \leq t_0 \\ 1 & \text{if} \quad \omega(x) \geq t_1 \\ R & \text{if} \quad t_0 < \omega(x) < t_1 \end{cases} \tag{2}$$

with the constraint $t_0 \leq t_1$. R represents the rejection of the example x. Figure 1 shows the distribution of the two classes on the classifier output. The two thresholds t_0 and t_1 divide the classifier output into three decision regions ($\{\Psi(x) = 1, \Psi(x) = 0, \Psi(x) = R\}$). The performance of the classifier depends on the following values: the error rate $E = p(\Psi(x) \neq y, \Psi(x) \neq R)$ (represented by the FP and FN areas in the figure 1), the penalty of an error λ_E, the accuracy $A = p(\Psi(x) = y)$ (represented by the TP and TN areas), the penalty of a good classification λ_A, the rejection rate $R = p(\Psi(x) = R)$ (represented by the R area) and the penalty of a rejection λ_R. Note that we have $A + R + E = 1$. The performance of a reject classifier is measured by its expected loss:

$$L(\Psi) = \lambda_A A + \lambda_E E + \lambda_R R \tag{3}$$

The objective is to find the thresholds t_0 and t_1 minimizing the expected loss of the classifier. For that we use the Chow's rule (Chow, 1970) that consider a Bayesian scenario where the output of the classifiers is the posterior probability of the positive class $\omega(x) = p(1|x)$. Let's the three loss functions L_1, L_0 and L_R that represent the expected loss that is obtained in assigning an example x to respectively the class 1, 0 or R.

$$\begin{aligned} L_1(x) &= \lambda_A \omega(x) + \lambda_E (1 - \omega(x))) \\ L_0(x) &= \lambda_E \omega(x) + \lambda_A (1 - \omega(x)) \\ L_R(x) &= \lambda_R \end{aligned} \tag{4}$$

From these formulas we can compute directly the optimal decision thresholds in solving the equations

$$\begin{aligned} \frac{L_1(x)}{L_R(x)} = 1 \quad &\Rightarrow \quad t_1^* = \frac{\lambda_R - \lambda_E}{\lambda_A - \lambda_E} \\ \frac{L_0(x)}{L_R(x)} = 1 \quad &\Rightarrow \quad t_0^* = \frac{\lambda_R - \lambda_A}{\lambda_E - \lambda_A} \end{aligned} \tag{5}$$

3.3 Cascade of Reject Classifiers

Our cascade system is a sequence of D classifiers with reject option $\Psi_1, \ldots, \Psi_D$ of increasing cost, illustrated by the figure 2. The i-th classifier Ψ_i receives all examples rejected by the classifier Ψ_{i-1}, makes predictions and sends all rejected examples to Ψ_{i+1}. The last classifier Ψ_D has no reject option and makes a prediction for all received examples. The first classifier Ψ_1 receives all examples. For the moment, we consider that the order of the variables is fixed, the classifier Ψ_i uses only the i first features so its cost is $\sum_{j=1}^{i} c_j$. For each classifier Ψ_i, its error rate E_i, accuracy A_i and rejection rate R_i are computed as :

$$\begin{aligned} E_i &= p(\Psi_i(x) \neq y, \Psi_i(x) \neq R | \Psi_j(x) = R \; \forall j \in [1, i-1]) \\ A_i &= p(\Psi_i(x) = y | \Psi_j(x) = R \; \forall j \in [1, i-1]) \\ R_i &= p(\Psi_i(x) = R | \Psi_j(x) = R \; \forall j \in [1, i-1]) \end{aligned} \tag{6}$$

From these formulas, we can define the loss L_i of each classifier of the cascade by a weighted combination of their error rate, accuracy and rejection rate. The weight of a good classification is the cost of the used variables, the weight of an error is the cost of

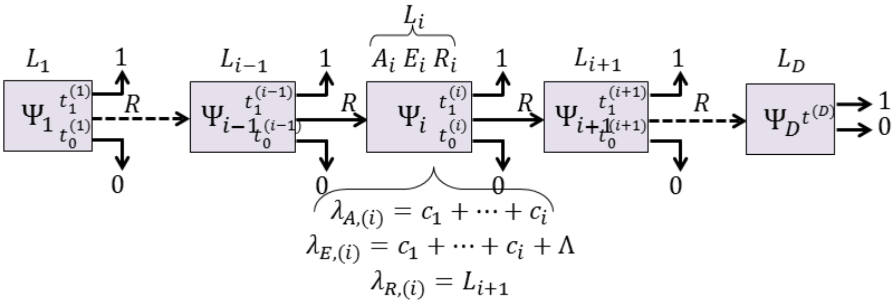


Figure 2: Cascade of D reject classifiers.

the used variables plus the penalty of misclassification. When an example is rejected, it is sent to the next classifier so the weight of rejection is the loss of the next classifier L_{i+1}. The loss of an entire cascade L can be computed recursively by:

$$\begin{aligned} L &= L_1 \\ L_i &= A_i \sum_{j=1}^{i} c_j + E_i\left(\sum_{j=1}^{i} c_j + \Lambda\right) + R_i L_{i+1} \\ L_D &= A_D \sum_{j=1}^{D} c_j + E_D\left(\sum_{j=1}^{D} c_j + \Lambda\right) \end{aligned} \tag{7}$$

The optimization of the cascade consist of finding the optimal rejection areas of each classifier that minimize the loss of the cascade. For each classifier of the cascade, the rejection area i.e. the thresholds t_0 and t_1, can be computed in using the Chow's rule. For the classifier Ψ_i the penalty of a good classification is $\lambda_A = \sum_{j=1}^{i} c_j$, the penalty of an error is $\lambda_E = \sum_{j=1}^{i} c_j + \Lambda$ and the penalty of a rejection is $\lambda_R = L_{i+1}$. In using the formulas (5) we obtain the optimal rejection area of the classifier Ψ_i.

$$t^*_{0,(i)} = \frac{L_{i+1} - \sum_{j=1}^{i} c_j}{\Lambda} \qquad t^*_{1,(i)} = \frac{\sum_{j=1}^{i} c_j + \Lambda - L_{i+1}}{\Lambda} \tag{8}$$

Unfortunately, we can not simply use these formulas on each classifier to obtain the optimal cascade. The problem is that the classifiers and their performances are depending each other. When a new rejection area of a classifier is computed, the sets of examples rejected to the next classifiers change, the performances of the next classifiers and their penalties of rejection change too. A new rejection area has, therefore, to be computed. All rejection areas, performances and penalties of all classifiers are circularly dependent. To solve this optimization problem we propose a heuristic described in the algorithm 1. The cascade is initialized as the basic classifier i.e. all classifiers reject all examples and all examples are sent to the last classifier using all variables. The iterative procedure contains three steps. The first one is to compute the accuracy, error rate and rejection rate of all classifiers. Then the penalties of rejection of all classifiers (excepted the last one) are computed in using the formula (7). The penalty of rejection depends on the performances of the next classifier, the penalties are therefore computed from the classifier Ψ_{D-1} to the classifier Ψ_1. Finally, the two rejection thresholds are computed for each classifier from the penalties of good classification, misclassification, and rejection. This procedure is iterated $MaxIter$ times, $MaxIter$ is a parameter to be chosen by the users. In the results section, we investigate empirically the impact of this parameter and select $MaxIter = 10$.

3.4 Order of the Variables

In the previous section, we considered that the order of the variables in the cascade was fixed, but in real case the variable order is rarely known. The performance of the cascade depend highly on the order, we want the most informative and less expensive variables at the beginning and the less informative and most expensive at the end. The usefulness of a variable is not correlated to its cost and is depending on the variable selected in the previous classifiers. For these reasons, it is not easy to compute the quality of the variables and determine their position in the cascade. One solution is to test all orders and select the one that produce the best cascade. However, there are $D!$ possible orders, this method is intractable for $D > 10$. We propose a heuristic, in the algorithm 2, that selects an order of the variable. The heuristic begins with an empty set of variables and selects one by one each variable. At each iteration i, $i-1$ variables have been already selected and are used to construct a cascade of size $i-1$. All non-selected variables are tested to form the i-th stage of the cascade. We select the variable that minimizes the loss of the cascade. The procedure is iterated until all variables have been selected.

Algorithm 1: Optimization algorithm of the reject areas.

procedure REJECT AREAS OPTIMIZATION
  // Initialization
  for i from 1 to $D-1$ **do**
    $t_{0,(i)} \leftarrow 0$; $t_{1,(i)} \leftarrow 1$
  end for
  $t_{0,(D)} \leftarrow 0.5$; $t_{1,(D)} \leftarrow 0.5$
  for $nbiter$ from 1 to $MaxIter$ **do**
    // Computation of the reject classifiers performances
    $L \leftarrow newL$
    for i from 1 to D **do**
      $A_{(i)} \leftarrow accuracy\ of \Psi_i$
      $E_{(i)} \leftarrow error\ rate\ of \Psi_i$
      $R_{(i)} \leftarrow rejection\ rate\ of \Psi_i$
    end for
    // Computation of the rejection costs
    $\lambda_D \leftarrow 0$
    for i from $D-1$ to 1 **do**
      $\lambda_{R,(i)} \leftarrow A_{(i+1)} \sum_{j=1}^{i+1} c_j + E_{(i+1)}(\sum_{j=1}^{i+1} c_j + \Lambda) + R_{(i+1)}\lambda_{R,(i+1)}$
    end for
    // Computation of the thresholds
    for i from 1 to $D-1$ **do**
      $(t_{0,(i)}, t_{1,(i)})$ computed from the cost $\lambda_G = \sum_{j=1}^{i} c_j$, $\lambda_E = \sum_{j=1}^{i} c_j + \Lambda$ and $\lambda_R = \lambda_{R,(i)}$
    end for
  end for
  return $(t_{0,(i)}, t_{1,(i)})\ \forall i \in [1, D-1]$
end procedure

Algorithm 2: Selection of the variables order.

procedure REJECT AREAS OPTIMIZATION
  $V \leftarrow \{v_1, \ldots, v_D\}$
  $Order \leftarrow \emptyset$
  for j from 1 to D **do**
    $best.L \leftarrow \Lambda$
    for i from 1 to $D-j+1$ **do**
      $Tested.Order \leftarrow concat(Order, V[i])$
      Construction of the cascade from $Tested.Order$
      $L \leftarrow$ loss of the cascade
      if $L < best.L$ **then**
        $best.L \leftarrow L$, $best.V \leftarrow V[i]$
      end if
    end for
    $Order \leftarrow Order + best.V$
    $V \leftarrow V - best.V$
  end for
  return $Order$
end procedure

4 EXPERIMENTS AND RESULTS

4.1 Study Design and Datasets

We perform a set of experiments to investigate the performance of our cascade method. For these experiments, we use several real medical and genomic datasets. The first one is the pima dataset (Smith et al., 1988) whose the objective is to predict signs of diabetes of 768 patients based on eight clinical variables. We select this dataset because the costs of variables are provided, this information is very rare in the public datasets. The second one is the Wisconsin Diagnostic Breast Cancer (wdbc) datasets whose the objective is to differentiate the malignant tumors from benign tumors of 569 patients based on 30 medical variables. Since the costs of variables were no available, we randomly draw from a uniform distribution $U[0,1]$ the cost of the variables. The third one is the lung cancer dataset (Bhattacharjee, 2001) whose the objective is to identify the adenocarcinoma from the other type of tumor based on the several thousand of gene expression. The last one is the prostate cancer dataset (Singh et al., 2002) whose the objective is to diagnosis cancer from safe tissues based on the several thousand of gene expression of 339 patients. For the two last datasets, since all gene expressions have been measured simultaneously with microarrays, the costs of all variables are equal. For all datasets, we normalize the costs of the variables such that the sum of all costs is 1. The basic classifier using all variables pays, therefore, one for each example.

We tested our method with two classification algorithms: the linear discriminant analysis (LDA) and the support vector machine (SVM) with a radial kernel. For high dimensional datasets, like the lung and prostate cancer dataset, a feature selection step is included in the classification in order to reduce the number of variables. We have used a filter method based on the t-test score to select the best variables. Note that the variable selection is performed in the classifier construction in order to avoid any selection bias (Ambroise and McLachlan, 2002).

The objective of our method is to reduce the classification cost and have a lower loss than the basic classifier. We, therefore, compare the performances of our cascade of the performance of the basic classifiers. One of the key points of the cascade construction is the selection of the variables order for which we have proposed a heuristic (algo 2). In order to show the usefulness of our heuristic, we compare our method to the performance of cost based order cascade. In the cost-based order cascade, the variables are ordered by their increasing cost. The cascade be-

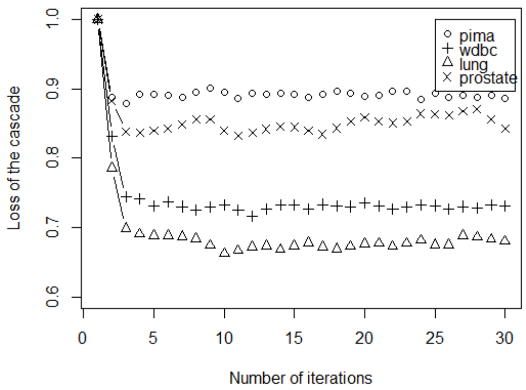


Figure 3: Loss of the cascade during the computation of the rejection areas.

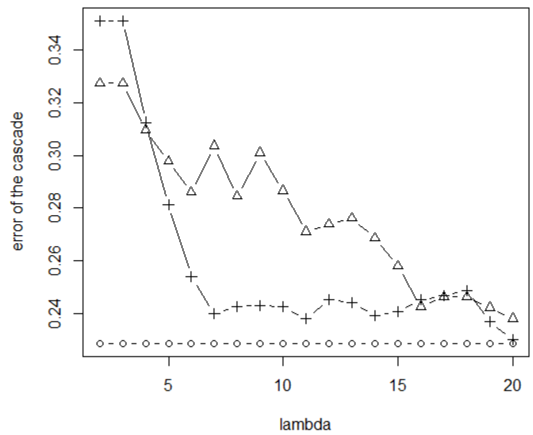
gins with the cheapest variables and finishes with the more expensive.

4.2 Sensitivity Analysis

Our method depends on two parameters: the number of iterations in the heuristic of rejection areas computation $MAXITER$ and the penalty of an error Λ. We investigate the impact of these two parameters on the behavior of our method.

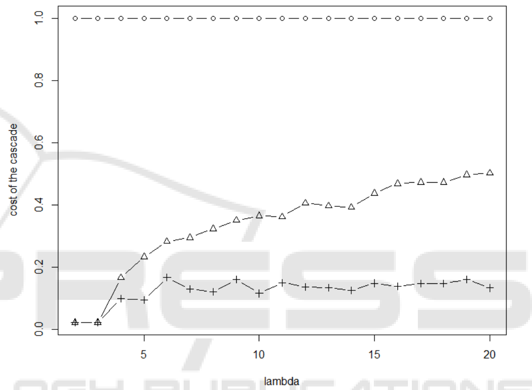
The figure 3 shows the loss of the cascade during the heuristic of the rejection areas computation for the four datasets with the LDA classifier. The loss values have been normalized such that all curves are plotted in the same graphics, we set the loss of the cascade at its initialization to 1. The figure shows that the cascade converges quickly toward a stable solution for all datasets. Moreover, we see that the loss of the solution is much lower than the loss of initialization. This last point shows that our heuristic provides good rejection areas for the cascade. According to these results, we choose to set $MAXITER = 10$, this value is enough to reach a stable solution and limits the computation time of cascade learning.

The figures 4, 5 and 6 gives respectively the error rate vs λ, the cost of the cascade vs Λ and the cost vs the error rate with the pima dataset and the LDA classifier. We do not have the space to put the graphics of the other datasets and classifiers, but they are similar to these figures and lead to the same conclusions. The dot represents the basic classifier, the triangle line is the cost based order cascade and the cross line is the heuristic based order cascade. Λ is increasing with the cost of the cascade and decreasing with its error rate. Λ controls the trade-off between the error rate and the variable cost. For a low value of Λ, the misclassifications are more tolerated, fewer variables are therefore needed, but the error rate increases. At the extreme, $\Lambda \leq 2$ in these figures, the cascade keeps



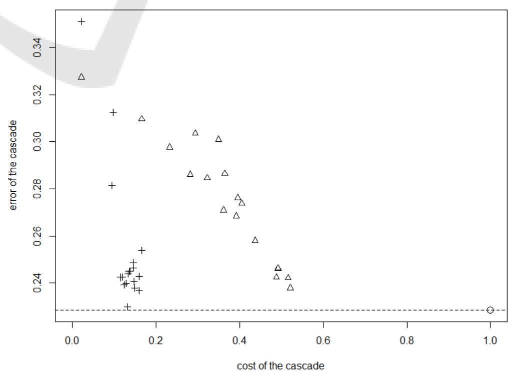
◯:basic classifier, △:cost based, +:heuristic based

Figure 4: Error rate of the cascade in function on Λ.



◯:basic classifier, △:cost based, +:heuristic based

Figure 5: Cost of the cascade in function on Λ.



◯:basic classifier, △:cost based, +:heuristic based

Figure 6: Error of the cascade in function on prediction cost. The performances are presented by a set of points because their are depending on the value of the parameter Λ.

only the first variable for all examples. For a high value of Λ, the misclassifications are very penalized, the cascade needs more variables in order to get more

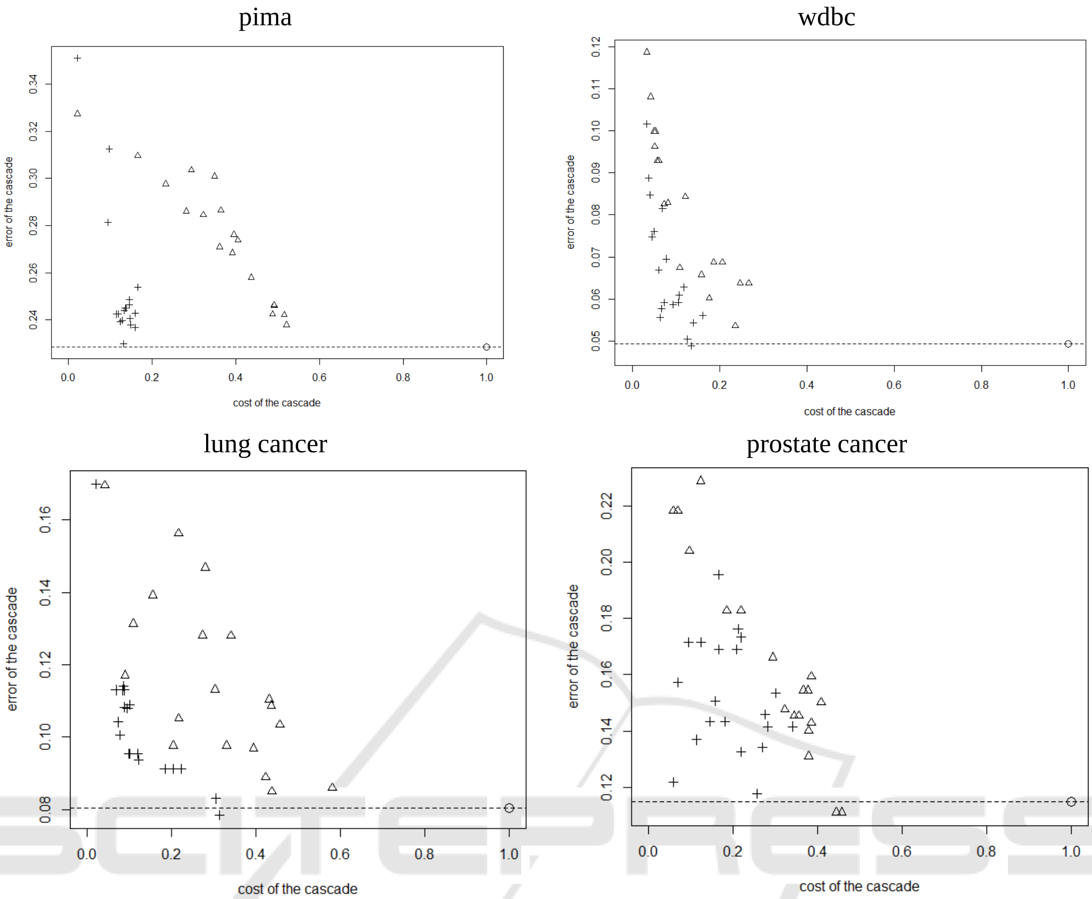


Figure 7: Classification cost vs Error rate plots for all datasets with linear discriminant analysis. The performances are presented by a set of points because their are depending on the value of the parameter Λ.

information and minimize the risk of error. We see that the error rate of the cascade is never lower than the error rate of the basic classifier. That is logic since the basic classifier uses all information, i.e. all variables for all examples. The error rate of the cascade can be only higher or equal than the error of the basic classifier. Note that there is always a value of Λ where the error rate of the cascade reaches the error rate of the basic classifier. In the figure 4, it is $\Lambda = 20$ for heuristic based order cascade and $\Lambda > 20$ for cost based order cascade. This point is interesting because it corresponds to a cascade that does not decrease the accuracy of the classifier. Let focus on the behavior of the cost based order cascade (triangle curve) in the loss figure. For low values of Λ, the loss of cost based order cascade is the same than heuristic based order cascade, for high values of Λ it reaches the loss of basic classifier. The reason is that the cost based order cascade favors cheap variables, for low values of Λ the cost of the cascade is more important than its error rate, the cost based order cascade is therefore well adapted.

4.3 Classification Results

The figures 7 and 8 show the classification cost vs error rate plot for all datasets with the linear discriminant analysis. In these graphics, the closest a point is from the left bottom corner, the better the performance is. The dot represents the performance of the basic classifier, its cost is 1 by definition and its error rate is represented by the dotted line. The triangles and the crosses represent respectively the performance of the cost based order cascade and the heuristic based order cascade. The performances are presented by a set of points because their are depending on the value of the parameter Λ. In all graphics, we see that the cascade can decreases strongly the cost of the classification. With the same accuracy as the basic classifier, we can reduce the cost by 85% for pima datasets, 86% for wdbc dataset, 70% for the lung cancer dataset and 74% for prostate cancer datatset with the LDA classifier and by 74% for pima datasets, 73% for wdbc dataset, 90% for the lung cancer dataset and 80% for prostate cancer datatset with the SVM classi-

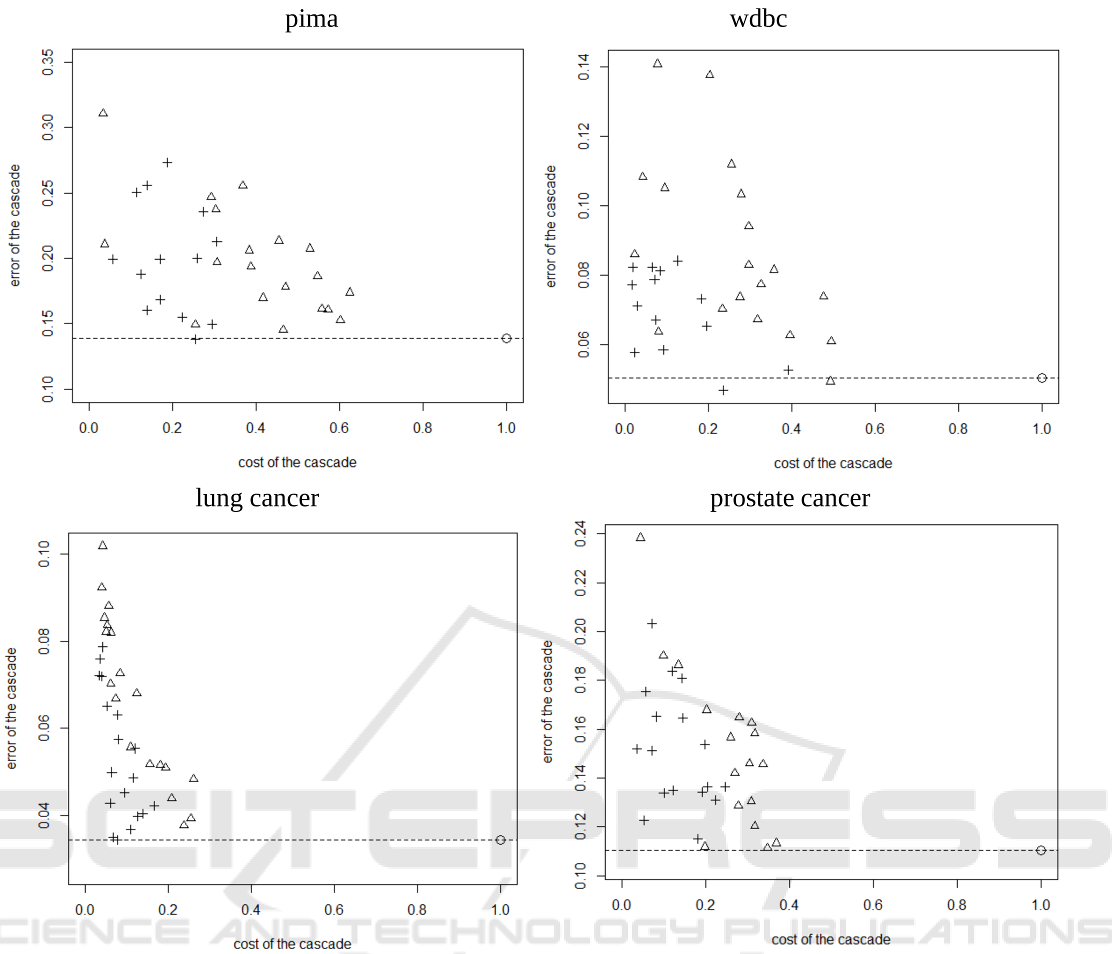


Figure 8: Classification cost vs Error rate plots for all datasets with support vector machine. The performances are presented by a set of points because their are depending on the value of the parameter Λ.

fier. This cost can again be decreased, if we accept to increase the error rate. We also see that in all graphics that the triangles clearly dominate the crosses. This means that the heuristic based order cascade outperforms the cost based order cascade.

5 CONCLUSIONS

The cascade methods are very promising for personalized medicine since the prediction system and its cost are adapted to each patient. There are some problems. The first one is the problem of high dimension data like the omics data. If the number of variables is very high (several thousand and more) the heuristic of order selection is computationally intractable. In the current work, we deal with this problem in making a classic variable selection step before the construction of the cascade. A more efficient solution would be to perform the selection in taking account of the variable cost and during the construction of the cascade. The second question is the problem of cost based problems. In the current work the costs are unique and fixed for each variable. In another context they may have several costs, for example, the medical exam to obtain some variables may have a cost in money, duration and a risk of secondary effect. All these costs impact the performance of the cascade. They may also have interactions between the costs. The cost of a variable v_i can decrease if a variable v_j has already been measured. We will study these new interesting problems on future works.

REFERENCES

Ambroise, C. and McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarray gene expression data. *Proc. Natl. Acad. Sci.*, 99(10):6562–6566.

Bhattacharjee, A. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *PNAS*, 98(24):13790–5.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

Diaz-Uriarte, R. and Alvarez de Andres, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3).

Dudoit, S., Fridlyand, J., and Speed, P. (2002). Comparison of discrimination methods for classification of tumors using gene expression data. *Journal of American Statististial Association*, 97:77–87.

Furey, T., Cristianini, N., Duffy, N., Bednarski, D., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914.

Hood, L. and Friend, S. H. (2011). Predictive, personalized, preventive, participatory (p4) cancer medicine. *Nat Rev Clin Oncol*, 8(3):184–187.

Kapoor, A. and Horvitz, E. (2009). Breaking boundaries: Active information acquisition across learning and diagnosis. *Advances in neural information processing systems*.

Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwarb, M., Antonescu, C., Peterson, C., and Meltzer, P. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medecine*, 7:673–679.

Nan, F., Wang, J., and Saligrama, V. (2015). Feature-budgeted random forest. *International Conference on Machine Learning*.

Raykar, V. C., Krishnapuram, B., and Yu, S. (2010). Designing efficient cascaded classifiers: tradeoff between accuracy and cost. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 853–860. ACM.

Saar-Tsechansky, M., Melville, P., and Provost, F. (2009). Active feature-value acquisition. *Management Science*, 55(4):664–684.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., and Ladd, C. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell.*, 1(2):203–209.

Smith, J. W., Everhart, J., Dickson, W., Knowler, W., and Johannes, R. (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association.

Tan, Y. F. and yen Kan, M. (2010). Cost-sensitive attribute value acquisition for support vector machines. Technical report, National University of Singapore.

Trapeznikov, K. and Saligrama, V. (2013). Supervised sequential classification under budget constraints. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 581–589.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, 57(2):137–154.

Wang, L., Lin, J., and Metzler, D. (2011). A cascade ranking model for efficient ranked retrieval. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 105–114, New York, NY, USA. ACM.

Yang Pengyi; Hwa Yang Yee; Bing B. Zhou;, B. B. Z. (2010). A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296.