Cangelosi et al. BMC Bioinformatics 2014, 15(Suppl 5):S4
http://www.biomedcentral.com/1471-2105/15/S5/S4

BMC Bioinformatics

RESEARCH Open Access

Use of Attribute Driven Incremental Discretization and Logic Learning Machine to build a prognostic classifier for neuroblastoma patients

Davide Cangelosi[1], Marco Muselli[2†], Stefano Parodi[2], Fabiola Blengio[1], Pamela Becherini[1], Rogier Versteeg[3], Massimo Conte[4], Luigi Varesio[1*†]

From Tenth Annual Meeting of the Italian Society of Bioinformatics (BITS)
Udine, Italy. 21-23 May 2013

Abstract

Background: Cancer patient's outcome is written, in part, in the gene expression profile of the tumor. We previously identified a 62-probe sets signature (NB-hypo) to identify tissue hypoxia in neuroblastoma tumors and showed that NB-hypo stratified neuroblastoma patients in good and poor outcome [1]. It was important to develop a prognostic classifier to cluster patients into risk groups benefiting of defined therapeutic approaches. Novel classification and data discretization approaches can be instrumental for the generation of accurate predictors and robust tools for clinical decision support. We explored the application to gene expression data of Rulex, a novel software suite including the Attribute Driven Incremental Discretization technique for transforming continuous variables into simplified discrete ones and the Logic Learning Machine model for intelligible rule generation.

Results: We applied Rulex components to the problem of predicting the outcome of neuroblastoma patients on the bases of 62 probe sets NB-hypo gene expression signature. The resulting classifier consisted in 9 rules utilizing mainly two conditions of the relative expression of 11 probe sets. These rules were very effective predictors, as shown in an independent validation set, demonstrating the validity of the LLM algorithm applied to microarray data and patients' classification. The LLM performed as efficiently as Prediction Analysis of Microarray and Support Vector Machine, and outperformed other learning algorithms such as C4.5. Rulex carried out a feature selection by selecting a new signature (NB-hypo-II) of 11 probe sets that turned out to be the most relevant in predicting outcome among the 62 of the NB-hypo signature. Rules are easily interpretable as they involve only few conditions.
Furthermore, we demonstrate that the application of a weighted classification associated with the rules improves the classification of poorly represented classes.

Conclusions: Our findings provided evidence that the application of Rulex to the expression values of NB-hypo signature created a set of accurate, high quality, consistent and interpretable rules for the prediction of neuroblastoma patients' outcome. We identified the Rulex weighted classification as a flexible tool that can support clinical decisions. For these reasons, we consider Rulex to be a useful tool for cancer classification from microarray gene expression data.

* Correspondence: luigivaresio@ospedale-gaslini.ge.it
† Contributed equally
[1]Laboratory of Molecular Biology, Gaslini Institute, Largo Gaslini 5, 16147 Genoa, Italy
Full list of author information is available at the end of the article

Background

Neuroblastoma (NB) is the most common solid pediatric tumor, deriving from ganglionic lineage precursors of the sympathetic nervous system [2]. It shows notable heterogeneity of clinical behavior, ranging from rapid progression, associated with metastatic spread and poor clinical outcome, to spontaneous, or therapy-induced regression into benign ganglioneuroma. Age at diagnosis, stage and amplification of the N-myc proto-oncogene (*MYCN*) are clinical and molecular risk factors that the International Neuroblastoma Risk Group (INRG) utilized to classify patients into high, intermediate and low risk subgroups on which current therapeutic strategy is based. About fifty percent of high-risk patients die despite treatment making the exploration of new and more effective strategies for improving stratification mandatory [3].

The availability of genomic profiles improved our prognostic ability in many types of cancers [4]. Several groups used gene expression-based approaches to stratify NB patients. Prognostic gene signatures were described [5-11] and classifier proposed to predict the risk class and/or patients' outcome [5-13]. We and other scientific groups have identified tumor hypoxia as a critical component of neuroblastoma progression [14-16]. Hypoxia is a condition of low oxygen tension occurring in poorly vascularized areas of the tumor which has profound effects on cell growth, genotype selection, susceptibility to apoptosis, resistance to radio- and chemotherapy, tumor angiogenesis, epithelial to mesenchymal transition and propagation of cancer stem cells [17-20]. Hypoxia activates specific genes encoding angiogenic, metabolic and metastatic factors [18,21] and contributes to the acquisition of the tumor aggressive phenotype [18,22,23]. We have used gene expression profile to assess the hypoxic status of NB cells and we have derived a robust 62-probe sets NB hypoxia signature (NB-hypo) [14,24], which was found to be an independent risk factor for neuroblastoma patients [1].

The use of gene expression data for tumor classification is hindered by the intrinsic variability of the microarray data deriving from technical and biological variability. These limitation can be overcome by analyzing the results through algorithms capable to discretize the gene expression data in broad ranges of values rather than considering the absolute values of probe set expression. We will focus on the discretization approach to deal with gene expression data for patients' stratification in the present work.

Classification is central to the stratification of cancer patients into risk groups and several statistical and machine learning techniques have been proposed to deal with this issue [25]. We are interested in classification methods capable of constructing models described by a set of explicit rules for their immediate translation in the clinical setting and for their easily interpretability, consistency and robustness verification [15,26]. A rule is a statement in the form "if<premise> then<consequence>" where the premise is a logic product (AND) of conditions on the attributes of the problem and the consequence indicates the predicted output. Most used rule generation techniques belong to two broad paradigms: decision trees and methods based on Boolean function synthesis.

The decision tree approach implements discriminant policies where differences between output classes are the driver for the construction of the model. These algorithms divide iteratively the dataset into smaller subsets according to a divide and conquer strategy, giving rise to a tree structure from which an explicit set of rules can be easily retrieved. At each iteration a part of the training set is split into two or more subsets to obtain non-overlapping portions belonging to the same output class [27]. Decision tree methods provide simple rules, and require a reduced amount of computational resources. However, the accuracy of the models is often poor. The divide and conquer approach prevents the applicability of these models to relatively small datasets that would be progressively fractionated in very small, poorly indicative, subsets.

Methods based on Boolean function synthesis adopt an aggregative policy where some patterns belonging to the same output class are clustered to produce an explicit rule at any iteration. Suitable heuristic algorithms [28-30] are employed to generate rules exhibiting the highest covering and the lowest error; a tradeoff between these two objectives has been obtained by applying the Shadow Clustering (SC) technique [28] which leads to final models, called Logic Learning Machines (LLM), exhibiting good accuracy. The aggregative policy can also consider patterns already included in previously built rules; therefore, SC generally produces overlapping rules that characterize each output class better than the divide-and-conquer strategy. Clustering samples of the same kind permits to extract knowledge regarding similarities of the members of a given class rather than information on their differences. This is very useful in most applications and leads to models showing higher generalization ability, as shown by trials performed with SC [31,32].

LLM algorithms prevent the excessive fragmentation problem typical of divide-and-conquer approach but come at the expense of the need to implement an intelligent strategy for managing conflicts occurring when one instance is satisfied by more than one rules classifying opposite outcomes. LLM is a novel and efficient implementation of the Switching Neural Network (SNN) model [33] trained through an optimized version of the SC algorithm. LLM, SNN and SC have been successfully used in different applications: from reliability evaluation of complex systems [34] to prediction of social phenomena

[35], form bulk electric assessment [36] to analysis of biomedical data [15,31,32,37,38].

The ability of generating models described by explicit rules has several advantages in extracting important knowledge from available data. Identification of prognostic factors in tumor diseases [15,37] as well as selection of relevant features in microarray experiments [31] are only two of the valuable targets achieved through the application of LLM and SNN. In this analysis, to improve the accuracy of the model generated by LLM, a recent innovative preprocessing method, called Attribute Driven Incremental Discretization (ADID) [39] has been employed. ADID is an efficient data discretization algorithm capable of transforming continuous attributes into discrete ones by inserting a collection of separation points (cutoffs) for each variable. The core of ADID consists in an incremental algorithm that adds the cutoff iteratively obtaining the highest value of a quality measure based on the capability of separating patterns of different classes. Smart updating procedures enable ADID to efficiently get a (sub) optimal discretization. Usually, ADID produces a minimal set of cutoffs for separating all the patterns of different classes. ADID and LLM algorithms are implemented in Rulex 2.0 [40], a software suite developed and commercialized by Impara srl that has been utilized for the present work.

Blending the generalization and the feature selection strength of LLM and the efficiency of ADID in mapping continuous variables into a discrete domain with the stratification power of the NB-hypo signature we obtained an accurate predictor of NB patients' outcome and a robust tool for supporting clinical decisions. In the present work, we applied Rulex 2.0 components to the problem of classifying and predicting the outcome of neuroblastoma patients on the bases of hypoxia- specific gene expression data. We demonstrate that our approach generates an excellent discretization of gene expression data resulting in a classifier predicting NB patients' outcome. Furthermore, we show the flexibility of this approach, endowed with the ability to steer the outcome towards clinically oriented specific questions.

Results

Rulex model

We analyzed gene expression of 182 neuroblastoma tumors profiled by the Affymetrix platform. The characteristics of the NB patients are shown in Table 1 and are comparable to what previously described [6]. We selected this dataset because the gene expression profile of the primary tumor, performed by microarray, was available for each patient. "Good" or "poor" outcome are defined, from here on, as the patient's status "alive" or "dead" 5 years after diagnosis respectively.

We previously described the NB-hypo 62 probe sets signature that represents the hypoxic response of neuroblastoma cells [14,24] and used this signature for developing the hypoxia-based classifier to predict the patients' status utilizing ADID to convert the continuous probe sets values into discrete attributes and LLM algorithm to generate classification rules. Both techniques are implemented in Rulex 2.0. The first assessment of the classifier was done on the training set of 109 randomly chosen patients, while the remaining 73 patients were utilized to validate the predictions (Figure 1). The outcome of the classifier is a collection of rules, in the form if<premise> then <consequence>, where the premise includes conditions

Table 1 Characteristics of 182 neuroblastoma patients included in the study.

Risk factors and outcome		Training set (n = 109)[a]%		Independent test set (n = 73)[a]	
		Number	%	Number	%
Age at diagnosis (Years)	1	54	49	33	46
	> 1	55	51	40	54
INSS stage					
	1	29	26	14	19
	2	15	14	9	13
	3	16	15	7	9
	4	33	30	35	48
	4s	16	15	8	11
MYCN status					
	Normal	91	83	61	83
	Amplified	18	17	12	17
Outcome					
	Good	81	74	50	68
	Poor	28	26	23	32

[a] The number of patients is 109 in the training set and 73 in the test set. The data show the total number of patients and the relative percentage in each subdivision.

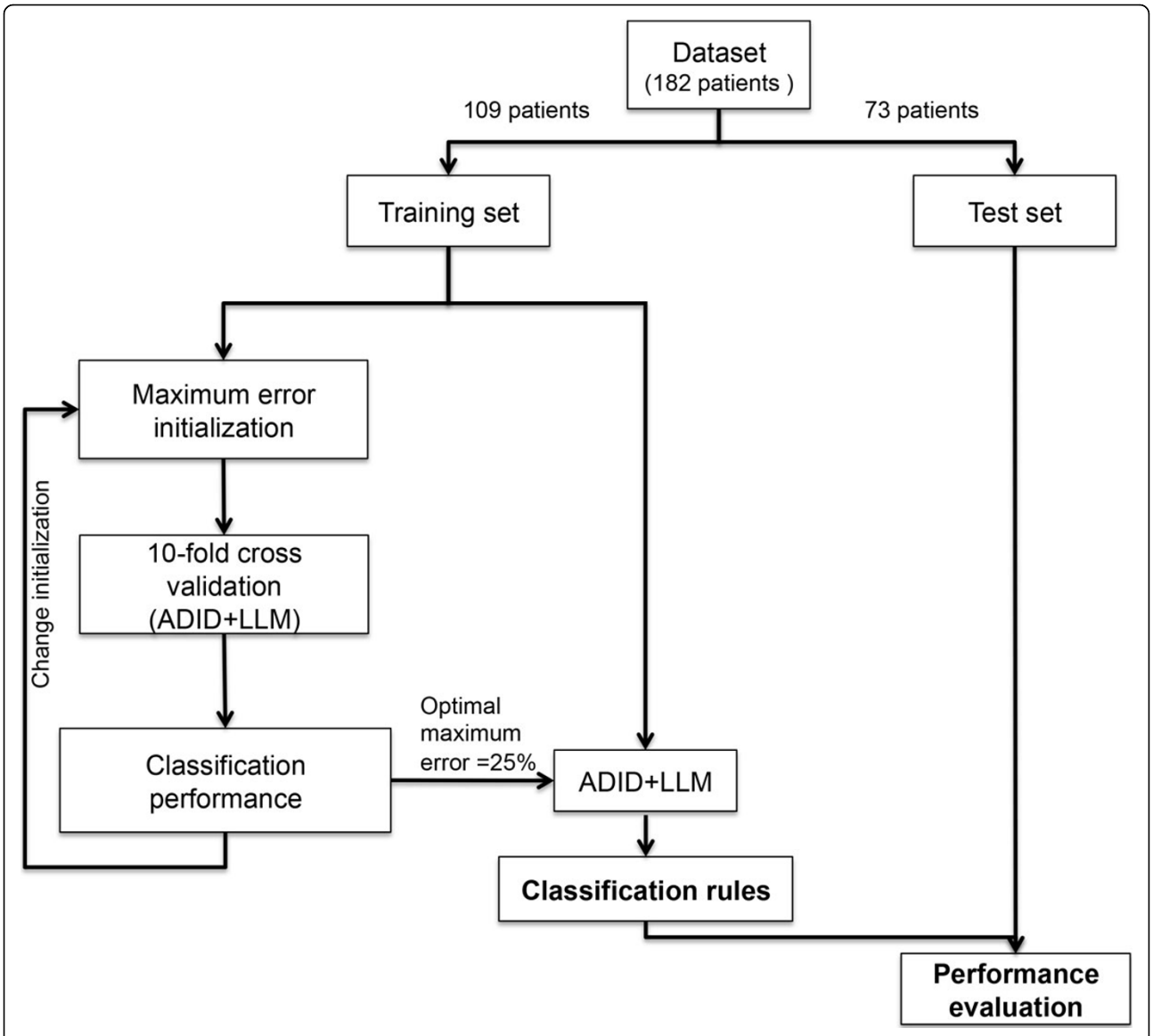

Figure 1 Rule generation workflow. The initial 182 patients dataset is randomly divided into training and test sets. The training set is used by the supervised learning procedure to iteratively calculate the LLM parameter:" maximum error allowed for a rule" by performing a complete 10-fold cross validation. The whole training set is randomly subdivided into 10 non-overlapping subsets, nine of which are used to train the classifier by employing ADID and LLM. The classifier is subsequently used to predict the outcome of the patients in the excluded subset. This procedure is repeated 10 times until every subset is classified once. Each parameter value is then evaluated according to the mean classification accuracy obtained in the cross validation. The parameter value, which obtained the highest mean accuracy, is selected to generate the final optimal classification rules. The rules are then tested on an independent cohort to assess their ability to predict patients' outcome.

based on the probe sets values and the consequence is the patient status. Rulex 2.0 will use these rules collectively for outcome prediction on the validation set.

The generation of the classification rules requires a discretization step because this simplifies the selection of the cut-off values of the probe sets expression (Figure 1). The discretization yielded one cutoff value for each probe set which was sufficient for modeling the outcome. The use of a single cutoff dichotomizes the probe set attributes in low or high expression, drastically reducing the influence of the technical and biological variability present in the models associated with the absolute values of probe sets expression.

Furthermore, a test on the maximum error allowed for a rule was defined. The final classification rules were trained with the optimal value of 25% associated with the maximal mean accuracy of 87%, determined by 10 fold cross validation analysis (Figure 1). The procedure generated 9 rules,

numbered from 1 to 9 (Rule ID) in Table 2 and based on conditions containing high or low probe sets expression value (above or below the set threshold respectively). Rule premises are limited to two conditions with the exception of rule 7 that has only one condition. Six rules predict good outcome and 3 poor outcome; they will be considered together in scoring the class attribution of new patients of the validation set. This is the optimal scenario proposed by Rulex 2.0 to utilize the 62 probe sets NB-hypo signature for classifying patients' outcome.

Three parameters, shown in Table 2, estimate the quality of the rules: 1) covering, measuring the generality, 2) error, measuring the ambiguity and 3) Fisher's p-value measuring the significance of each rule. The statistical significance of each rule by Fisher's exact test was very high ($p< 0.001$) providing strong evidence of the excellent quality of the rules. The covering ranged among rules from 48% (rule 6) to 80% (rule 7) for good outcome classes and ranged from 57% (rule 9) to 92% (rule 7) for poor outcome. Error ranged from 3.5% (rule 1) to 14% (rules 2 and 5) for good outcome class and ranged from 7.4% (rule 9) to 17% (rule 7) for poor outcome class. These rules have interesting features that will be addressed in detail. The first consideration is that the overall covering of the rules classifying good and poor outcome adds up to 380% and 209%, respectively, indicating overlap among rules. This is a characteristic of the LLM method implementing an aggregative, rather than fragmentation, policy as illustrated in the Materials and Methods section. However, overlap among the rules can lead to a conflict if the probe sets values of a patient satisfies two or more rules predicting opposite outcomes. To investigate whether overlapping among our rules can be source of conflicts we plotted in Figure 2 each patient's membership to the nine rules. The plot clusters the rules classifying good and poor outcome and the patients belonging to good or poor outcome classes. The results show that each patient is covered by at least one rule. Overlaps exist and occur mainly among rules predicting the same outcome (Figure 2B and 2C), which do not lead to classification conflicts. However, 30 patients (27% of the dataset) were covered by rules pertaining to opposite outcomes and are represented by those present in quadrant 2A (19 patients) or 2D (11 patients). Rulex 2.0 overcomes this problem by employing for assigning a class to a new sample a fast procedure that evaluates all the rules satisfied by it and their covering, thus generating a consensus outcome to be assigned to the sample as detailed in the Materials and Methods section.

A second characteristic of the classification rules is that they include only 11 out of 62 probe sets of the original NB-hypo signature. The relationship among probe sets and rules is shown in Table 3. Rulex 2.0 operated a second feature selection on the original 62 probe sets optimized for functioning in a binary profile of low and high probe set expression and gave rise to a modified hypoxia signature that we name NB-hypo-II.

It was of interest to assess the relative importance of each probe set in the classification scheme leading to identification of poor and good outcome patients. Negative or positive values indicate low or high expression associated with the predicted outcome and the relative relevance is measured by the absolute value. Ranking the probe sets may be of relevance to pick the genes for further validation on an alternative platform. It is interesting to note that probe set 217356_s_at is the most relevant for the classification of good and poor outcome patients.

A third interesting feature of the rules is the dichotomization of the expression values of each probe set even if Rulex 2.0 had no constraints on the discretization that could lead to multiple cutoffs or overlapping expression values in different rules. We were intrigued by these

Table 2 Classification rules.

Rule ID[a]	Cond 1		Cond 2		Predicted Outcome	Covering[b] (%)	Error[c] (%)	Fisher pvalue[d]
1	*IF(*	217356_s_at ≤ 721	226452_at < 326	*)THEN*	Good	80	3.5	<0.001
2	*IF (*	206686_at ≤ 26	226452_at ≤326	*)THEN*	Good	70	14	<0.001
3	*IF (*	200738_s_at ≤ 1846	230630_at> 23	*)THEN*	Good	62	10	<0.001
4	*IF(*	209446_s_at ≤ 57	223172_s_at < 73	*)THEN*	Good	60	10	<0.001
5	*IF(*	202022_at > 131	223193_x_at < 324	*)THEN*	Good	60	14	<0.001
6	*IF(*	224314_s_at ≤ 29	236180_at <13	*)THEN*	Good	48	7.1	<0.001
7	*IF(*	217356_s_at > 721		*)THEN*	Poor	92	17	<0.001
8	*IF(*	223172_s_at >73	226452_at> 326	*)THEN*	Poor	60	8.6	<0.001
9	*IF(*	206686_at > 26	223172_s_at>73	*)THEN*	Poor	57	7.4	<0.001

[a] Cond 1 and Cond 2 indicate the conditions into the premises of the rules.

[b] The covering accounts for the fraction of patients that verify the rule and belong to the target outcome.

[c] The error accounts for the fraction of patients that satisfy the rule and do not belong to the target outcome.

[d] Fisher p-value quantifies the statistical significance of the rule on the basis of the number of patients correctly and incorrectly classified by a rule and the number of patients of the dataset belonging to each specific outcome.

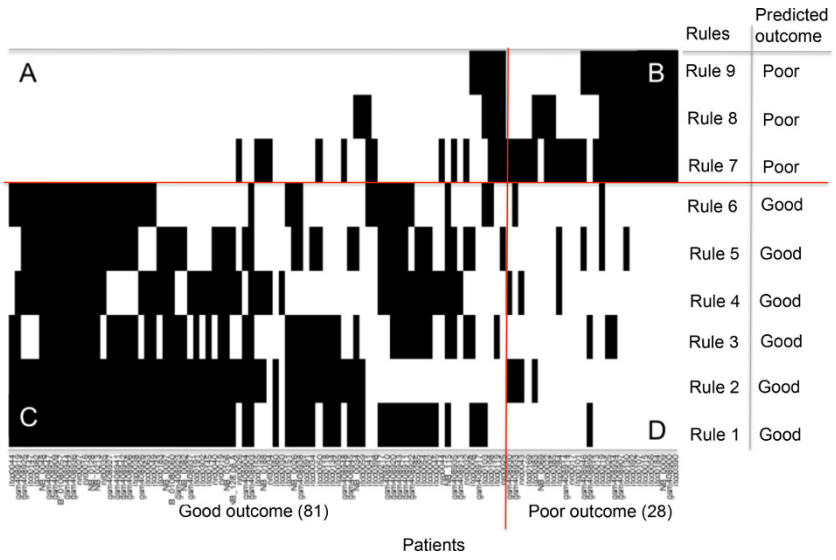

Figure 2 Patients representation in the rules of Table 2. Plot of the membership of the 109 patients (x axis) to the rules (y axis) of the classifier in Table 2. The rule identifier and predicted outcome are listed in right side of the plot. Two oriented lines divide the plot. The horizontal line separates the rules classifying good outcome from those classifying poor outcome. The vertical line separates good outcome from poor outcome patients. Each point of the plot indicates the membership of a patient to one rule. The two lines separate the plot in four sections labeled as A, B, C and D. Section A includes all the patients incorrectly classified by a poor outcome rule. Section B includes all the patients correctly classified by a poor outcome rule. Section C includes all the patients correctly classified by a good outcome rule and Section D includes all the patients incorrectly classified by a good outcome rule.

results and decided to examine the relationship between the cutoff expression values identified by Rulex and those obtained by Kaplan-Meyer analysis. The Kaplan-Meier algorithm calculates all possible cut- off points of a given probe set in a cohort of patients and selects the one maximizing the distinction between good and poor outcome. The results, shown in Table 4, demonstrated a general quite good concordance between the cutoff values of the Kaplan-Meier and those generated by Rulex. In particular, the two measures always differed by less than ±50% in magnitude, but only in some cases (probe sets 2, 3, 4, 5, and 8 for both overall and relapse free survival, and probe sets 6 and 7 for overall survival such a difference was lower than 25%. These (even if rather small) discrepancies are probably related to the capability of ADID to exploit the complex multivariate correlation among probe sets. Furthermore, we found a concordance also in the relationship between high/low probe set and outcome in Rulex derived rules and Kaplan-Meier plots (Table 4).

Table 3 Probe sets characteristics of the new NB-hypo-II signature.

Probe sets[a]	Rule ID[b]	Probe set relevance[c]	
		Good outcome	Poor outcome
200738_s_at	3	-0.49	0
202022_at	5	0.5	0
206686_at	2, 9	-0.48	0.26
209446_s_at	4	-0.25	0
217356_s_at	1, 7	-0.74	0.92
223172_s_at	4, 8, 9	-0.35	0.48
223193_x_at	5	-0.1	0
224314_s_at	6	-0.22	0
226452_at	l, 2, 8	-0.26	0.34
230630_at	3	0.13	0
236180_at	6	-0.25	0

[a] Affymetrix probe sets belonging to NB-hypo-II signature.

[b] Rule ID indicates the ID of the rules in which the probe set occurs (as illustrated in Table 2).

[c] Relevance measures the importance of the features included into the rules. The relevance calculated for each outcome is shown.

Outcome prediction

The ability of the rules in Table 2 to predict patients' outcome was tested on a 73 patients' independent dataset. Results are expressed as accuracy, recall and precision, assessing the performance in classifying good outcome, specificity and negative predictive values (NPV) assessing the performance in classifying poor outcome patients. The direct evaluation of the performance of the 9 rules on the validation set is represented by ADID in Table 5, by ADID +LLM in Table 6, the base configuration of Table 7 and by no dataset modification in Additional File 1. The results demonstrate a good accuracy comparable to what reported by other algorithms [13]. Furthermore, the classification of good outcome was superior to that of poor outcome

Table 4 Expression cut-offs from the Kaplan-Meier and from the rules.

Probe set ID[a]	NB-hypo-II[b]	Overall[c]				Relapse free[d]			
		Expression cut-off[e]		Worse[f]		Expression cut-off[e]		Worse[f]	
		Kaplan-Meier	Rulex	Kaplan-Meier	Rulex	Kaplan-Meier	Rulex	Kaplan-Meier	Rulex
1	223172_s_at	107	73	high	high	107	73	high	high
2	200738_s_at	1553	1846	high	high	1553	1846	high	high
3	209446_s_at	69	57	high	high	69	57	high	high
4	226452_at	280	326	high	high	280	326	high	high
5	217356_s_at	706	721	high	high	706	721	high	high
6	236180_at	18	13	hgih	hgih	13	13	high	hgih
7	202022_at	101	131	low	low	138	131	low	low
8	224314_s_at	25	29	high	high	25	29	high	high
9	206686_at	36	26	high	high	36	26	high	high
10	223193_x_at	495	324	high	high	572	324	high	high
11	230630_at	19	23	low	low	35	23	low	low

[a] Probe sets ID indicates the numerical identified of the probeset.

[b] NB-hyp-II indicates the list of probe sets of the NB-hypo signature belonging to the rules.

[c] Overall indicates the survival time between the time of an event or last follow up and the time of diagnosis.

[d] Relapse free indicates the survival time between the first relapse and the time of diagnosis.

[e] Expression cut-off indicates the optimal cut-off point of each probe set resulting from the Kaplan-Meier scan and from the rules.

[f] Worse indicates whether high or low expression of a given probe set is associated to the worse survival. Worse survival are calculated from the Kaplan-Meier curve or from the conditions included into the rules.

patients as shown for example by a recall of 90% relative to 57% of specificity.

PVCA analysis [41] was utilized to estimate the potential variability of experimental effects including batch. The analysis revealed that batch effect explained a moderate 21% of the overall variation in our dataset and a Frozen Surrogate Variable Analysis (FSVA) was employed for removing batch effect. The application of FSVA reduced batch effect to less than 0.05% of the total variation (data not shown).

We compared the performances achieved by ADID and LLM on the batch-adjusted dataset and those on the original dataset (no dataset modification) to measure the impact of batch effect on classifier performances. Performance obtained with the adjusted dataset turns out to be very similar to that obtained with original data as shown in Additional file 1 demonstrating that batch effect had negligible impact on the performances. Therefore, the dataset with no modifications was utilized for subsequent analysis.

We then compared the performance of the ADID discretization approach and those of commonly used discretization algorithms, namely: entropy based (EntMDL [42]), Modified Chi Square [43], ROC based (Highest Youden index([44]), and Equal frequency (i.e. median expression for each feature). Results detailed in Table 5 showed that

Table 5 Comparison among discretization algorithms' performance.

Algorithm[f]	Accuracy[a]	Recall[b]	Precision[c]	Specificity[d]	NPV[e]
ADID	80%	90%	82%	57%	72%
EntMDL	68%	60%	91%	87%	50%
Modified Chi2	71%	64%	91%	87%	53%
ROC-based	77%	84%	82%	61%	64%
Equal Frequency	68%	60%	91%	87%	50%

[a] Accuracy is the fraction of correctly classified patients and overall classified patients.

[b] Recall is the fraction of correctly classified good outcome patients and the overall predicted good outcome patients.

[c] Precision is the fraction of correctly classified good outcome patients and the predicted good outcome patients.

[d] Specificity is the fraction of correctly classified poor outcome patients and the overall poor outcome patients.

[e] NPV(negative predictive value) is the fraction of correctly classified poor outcome patients and the overall predicted poor outcome patients.

[f] Discretization algorithms utilized for comparison/

Table 6 Performance of classification algorithms.

Algorithm[f]	Accuracy[a]	Recall[b]	Precision[c]	Specificity[d]	NPV[e]
ADID + LLM	80%	90%	82%	57%	72%
Decision tree	63%	76%	72%	35%	40%
PAM	81%	90%	83%	61%	74%
SVM	84%	94%	84%	61%	82%

[a] Accuracy is the fraction of correctly classified patients and overall classified patients.

[b] Recall is the fraction of correctly classified good outcome patients and the overall predicted good outcome patients.

[c] Precision is the fraction of correctly classified good outcome patients and the predicted good outcome patients.

[d] Specificity is the fraction of correctly classified poor outcome patients and the overall poor outcome patients.

[e] NPV(negative predictive value) is the fraction of correctly classified poor outcome patients and the overall predicted poor outcome patients.

[f] Machine learning algorithms utilized for comparison.

Table 7 Performance comparison among the configurations in the weighted classification on the test set.

Configuration [f]	Accuracy[a]	Recall[b]	Precision[c]	Specificity[d]	Negative Predictive Value[e]
Base (not weighted)	80%	90%	82%	57%	72%
W26_74 (balanced outcome)	65%	63%	82%	70%	47%
Wl_1000 (bias poor)	66%	60%	86%	78%	47%
W1000_1 (bias good)	78%	98%	77%	35%	89%

[a] Accuracy is the fraction of correctly classified patients and overall classified patients.

[b] Recall is the fraction of correctly classified good outcome patients and the overall predicted good outcome patients.

[c] Precision is the fraction of correctly classified good outcome patients and the predicted good outcome patients.

[d] Specificity is the fraction of correctly classified poor outcome patients and the overall poor outcome patients.

[e] Negative predictive value is the fraction of correctly classified poor outcome patients and the overall predicted poor outcome patients.

[f] Configuration indicates the specific weights assigned to the outcomes in the weighted classification.

the discretization performed by ADID produced better accuracy with respect to the others (80% vs. 68%-77%).

We also compared the performance of LLM and those of Decision Tree [45], Support Vector Machines (SVM) [46], and Prediction Analysis of Microarrays (PAM) [47], to evaluate the ability of LLM to predict patients' outcome with respect to other standard supervised learning methods. Results in Table 6 revealed that ADID and LLM were able to predict patients' outcome with better performances with respect to the decision tree classifier. The performances of LLM, PAM and SVM classifiers were comparable.

Overall, the performance was good but unbalanced datasets tend to bias the performance towards the most represented class [48]. In our dataset the good outcome patients were more frequent (26 % poor and 74% good outcome). Therefore, we explored the possibility of utilizing a weighted classification system (WCS) [48-53] to improve the classification of poorly represented classes or to force the algorithm to maximize the performance on selected outcomes. The performance of the base configuration was taken as reference.

We performed a weighted classification accounting for the unbalanced class representation in the dataset (Table 7 configuration W26_74). The weight was calculated as the inverse proportion of the number of patients belonging to each class, about 3 times more weight on the poor outcome class. The major improvement over the base configuration was the specificity whereas all other parameters were similar or worst. Configuration W1_1000, was similar to W26_74, but set the weight of poor outcome 1000 times higher than that of the good one. Interestingly, its performance was very close to that of configuration W26_74 despite the disparity in the weight applied, suggesting that small changes in the relative weight of poor outcome are sufficient to optimize the results. In conclusion, increased weight on poor outcome augmented the percentage of correctly classified poor outcome patients even though a smaller number of patients were included in this class. This correction may be relevant when maximization of the specificity isof primary importance as in the case of using a prudent therapy.

In contrast, configuration W1000_1 sets the weight of good outcome 1000 times higher than that of poor outcome. The performance parameters were similar or higher than the base configuration with the exception of specificity that was quite low, a situation that appears symmetrical to those observed previously. The recall is almost absolute indicating the exceptional ability to classify good outcome. The drawback of this configuration is a very low percentage of correct poor outcome classification that is 35% of all poor outcome patients. This configuration may be useful in the case of using an aggressive therapy.

In conclusion, WCS can improve performance parameters of classification of poor or good outcome patients and may be particularly relevant in a situation where the dataset contains a major unbalance between classes and/or when clinical decisions may require minimizing false positives or false negatives.

Discussion

Our study is based on gene expression data derived through the analysis of primary neuroblastoma tumors by microarray on the Affymetrix platform. We focused on the expression of 62 probe sets comprising the NB-hypo signature that we have previously shown to represent the hypoxia status of neuroblastoma cells [24]. The association of hypoxia with poor prognosis in neuroblastoma patients was previously demonstrated [16,54]. We studied a cohort of 182 NB patients characterized by clinical and molecular data addressing the question of the potential prognostic value of this signature. Rulex 2.0 suite was used to train a model on a set of patients and validate it on an independent one. We demonstrate that Attribute Driven Incremental Discretization and Logic Learning Machine algorithms, implemented in Rulex 2.0, generated a robust set of rules predicting outcome of neuroblastoma patients using expression values of 11 probe sets, specific for hypoxia extracted from the gene NB-hypo expression profile.

Outcome prediction of NB patients was reported by several groups using a combination of different risk factors and utilizing various algorithms [5,6,12,55,56]. Several groups have used gene expression-based approaches to stratify neuroblastoma patients and prognostic gene signatures have been described often based on the absolute values of the probe sets after appropriate normalization [5-11,13]. Affymetrix GeneChip microarrays are the most widely used high-throughput technology to measure gene expression, and a wide variety of preprocessing methods have been developed to transform probe intensities reported by a microarray scanner into gene expression estimates [57]. However, variations from one experiment to another [58] may increase data variability and complicate the interpretation of expression analysis based on absolute gene expression values. We addressed the problem by applying the Attribute Driven Incremental Discretization algorithm [39] that maps continuous gene expression values into discrete attributes. Interestingly, the algorithm applied to our dataset showed that the introduction of a single cutoff was sufficient to create two expression patterns, operationally defined as low and high, capable of describing the probe set status accurately enough for effective patients classification. This approach minimizes the variability and errors associated with the use of absolute values to interpret microarray gene expression data.

The validity of the discretization implemented by Rulex 2.0 was further documented by an empirical validation where we calculated the optimal cutoff value for each of the 11 probe sets tested in a Kaplan-Meier analysis of the patients' survival. It is noteworthy that such analysis utilized the survival time of the patient as opposed to 5 years survival considered by Rulex 2.0. Nevertheless, we demonstrated that the cutoff values calculated by either approaches were rather similar, thus supporting the robustness of the ADID algorithm to identify relevant discrete groups of expression values. From a technical point of view ADID is a multivariate method searching for the minimum number of cutoffs that separate patients belonging to different classes. On the other hand, the Kaplan-Meier scan is a univariate technique having the aim of identifying the value of the probe set that maximizes the distance among the survival times of resulting groups. It should be noted that these two approaches are independent from each other since they are based on different algorithms and different classifications.

Only 11 out of 62 probe sets of the original signature were considered by LLM for building the classifier. This selection has a biological meaning. In fact, the original 62 probe sets NB-hypo signature was obtained following a biology driven approach [1] in which the prior knowledge on tumor hypoxia was the bases for the analysis and the signature was derived from hypoxic neuroblastoma cell lines [14]. Hence, NB-hypo is optimized for detecting tumor hypoxia. The importance of hypoxia in conditioning tumor aggressiveness is documented by an extensive literature [17,19,20,22,59]. However, NB-hypo was not optimized to predict outcome that is dependent on factors other than hypoxia. Rulex 2.0 performed a feature selection by identifying the 11 probe sets that were the most relevant in predicting outcome among those of the NB-hypo signature.

One key feature of LLM is to implement an aggregative policy leading to the situation in which one patient can be covered by more than one rule. This leads to the advantage of avoiding dataset fragmentation typical of the divide-and-conquer paradigm. Furthermore, the robustness of the resulting model is increased; in fact, if a patient satisfies more than one rule for the same output class, the probability of a correct classification is higher.

The same outcome is generally predicted by every rule verified by a given patient. However, there are situations in which a patient satisfies rules associated with opposite outcomes, thus generating a potential conflict. Rulex 2.0 overcomes this problem by adopting a procedure for assigning a specific class on the basis of the characteristics of the verified rules. A conflict should not necessarily be considered as a limit of the proposed approach, but it could reflect a source of ambiguity present in the dataset. If this were the case, any method building models from data would always reflect this ambiguity.

The generation of a predictive classifier based on gene expression obtained from different institutions raised the question of a possible batch effect in the data. We utilized Frozen Surrogate Variable Analysis method, a batch effect removal method capable of estimating the training batch, and used it as a reference for adjusting batch effect of other batches. In particular, those for which no information about the outcome is known. This was not possible with other known batch removal methods such as the Combating Batch Effects (Combat) [60], which adjusts the expression values of both training and test batch [61]. We compared the performances achieved by ADID and LLM on the batch-adjusted dataset and those on the original dataset (no dataset modification) to measure the impact of batch effect on classifier performances. Performance obtained with the adjusted dataset showed that batch effect had negligible impact on performances. Previous studies observed that the application of batch effect removal methods for prediction does not necessarily result in a positive or negative impact [61]. Furthermore, batch effect removal methods may remove the true biologically based signal [61]. For this reasons, the analysis of the present manuscript was performed on the original dataset excluding any modification for batch removal.

The 9 rules generated by ADID and LLM achieved a good accuracy on an independent validation set. We compared the accuracy of our new classifier with that of the

top performing classifiers for NB patients' outcome prediction listed in [13] to study the concordance of the performance achieved by the 9 rules with that of other previously published classifiers. The classifiers were generated on different signatures and algorithms and the accuracy reported ranged from 80% to 87%. Note that those classifiers were validated on a different test set; they utilized different algorithms and signatures. We concluded that the accuracy of our rules and that of other techniques reported in literature were concordant.

ADID represents an innovative discretization method that was used in combination with LLM for classification purposes. In the present study, ADID demonstrated to outperform other discretization algorithms, based on univariate analysis, indicating the capability of ADID to exploit the complex correlation structure commonly encountered in biomedical studies, including gene expression data sets. Moreover, ADID-LLM showed better performances with respect to the decision tree classifier. This was somewhat predictable because the lower performance of the decision tree with respect to other approaches has been pointed out in literature [15]. The LLM rules, SVM and PAM classifier achieved similar performances. The good performance achieved by ADID and LLM and the explicit representation of the knowledge extracted from the data provided by the rules demonstrated the utility of ADID and LLM in patients' outcome prediction.

Our dataset suffers of class-imbalance as many other datasets [48-53]. In fact, the good outcome class is overexpressed with respect to the poor outcome class. Rulex 2.0 implements a novel algorithmic strategy that allows setting up different weights to outcomes biasing the assignment to a class towards that of interest. It should be noted that weighting is effective only on class assignment for patients verified by conflicting rules. We have utilized the weight approach to represent the situation in which either poor or good outcome was favored or to address the imbalance between good and poor outcome patients in our dataset. We found that, in the absence of predefined weights, the algorithm generates a good performance somewhat unbalanced towards better classification of good outcome patients. By changing the weights we were in the position of steering the prediction towards a high precision in classifying poor outcome patients or in privileging the specificity. This tool may be or practical importance in the decision making process of clinicians that are confronted with difficult therapeutic choices.

Conclusions

We provided the first demonstration of the applicability of data discretization and rule generation methods implemented in Rulex 2.0 to the analysis of microarray data and generation of a prognostic classifier. Rulex automatically derived a new signature, NB-hypo II, which is instrumental in predicting the outcome of NB patients. The performances achieved by Rulex are comparable and in some case better than those of other known data discretization and classification methods. Furthermore, the easy interpretability of the rules and the possibility to employ weighted classification make Rulex 2.0 a flexible and useful tool to support clinical decisions and therapy assignment.

Methods

Patients

Affymetrix GeneChip HG-U133plus2.0 enrolled 182 neuroblastoma patients on the bases of the availability of gene expression profile. Eighty-eight patients were collected by the Academic Medical Center (AMC; Amsterdam, Netherlands) [1,62]; 21 patients were collected by the University Children's Hospital, Essen, Germany and were treated according to the German Neuroblastoma trials, either NB97 or NB2004; 51 patients were collected at Hiroshima University Hospital or affiliated hospitals and were treated according to the Japanese neuroblastoma protocols [63]; 22 patients were collected at Gaslini Institute (Genoa, Italy) and were treated according to Italian AIEOP protocols. The data are stored in the R2 microarray analysis and visualization platform (AMC and Essen patients) or at the BIT-neuroblastoma Biobank of the Gaslini Institute. The investigators who deposited data in the R2 repository agree to use them for this work. In addition, we utilized data present on the public database at the Gene Expression Omnibus number GSE16237 for Hiroshima patients [63]. Informed consent was obtained in accordance with institutional policies in use in each country. In every dataset, median follow-up was longer than 5 years, tumor stage was defined as stages 1, 2, 3, 4, or 4s according to the International Neuroblastoma Staging System (INSS), normal and amplified MYCN status were considered and two age groups were considered, those with age at diagnosis smaller than 12 months and greater or equal to 12 months. Good and poor outcome were defined as the patient's status alive or dead 5 years after diagnosis. The characteristics of the patients are shown in Table 1.

Batch effect measure and removal

The PVCA approach [41] was used to estimate the variability of experimental effects including batch. The pvca package implemented in R was utilized to perform the analysis setting up a pre- defined threshold of 60%. The analysis included Age at diagnosis, MYCN amplification, INSS stage, and Outcome and Institute variables. The estimation of experimental effects was performed before and after the batch effect removal.

The frozen surrogate variable analysis (FSVA) implemented in the sva package [64] was utilized for removing

the batch effect from the training and the test sets. The parametric prior method and the Institute batch variable were set up for the analysis.

Gene expression analysis

Gene expression profiles for the 182 tumors were obtained by microarray experiment using Affymetrix GeneChip HG-U133plus2.0 and the data were processed by MAS5.0 software according to Affymetrix guideline.

Preprocessing step

To describe the procedure adopted to discretize values assumed by the probe sets a basic notation must be introduced. In a *classification* problem d-dimensional examples $x \in X \subset \Re^d$, are to be assigned to one of q possible classes, labeled by the values of a categorical output y. Starting from a *training set* S including n pairs $(\boldsymbol{x}_i, y_i)$, $i = 1, ..., n$, deriving from previous observations, techniques for solving classification problems have the aim of generating a model $g(\boldsymbol{x})$, called *classifier*, that provides the correct answer $y = g(\boldsymbol{x})$ for most input patterns $\boldsymbol{x}$. Concerning the components x_j two different situations can be devised:

1. **ordered variables**: x_j varies within an interval $[a,b]$ of the real axis and an ordering relationship exists among its values.
2. **nominal (categorical) variables**: x_j can assume only the values contained in a finite set and there is no ordering relationship among them.

A discretization algorithm has the aim of deriving for each ordered variable x_j a (possibly empty) set of cutoffs γ_{jk}, with $k = 1, ..., t_j$, such that for every pair $\boldsymbol{x}_u$, $\boldsymbol{x}_v$ of input vectors in the training set belonging to different classes ($y_u \neq y_v$) their discretized counterparts $\boldsymbol{z}_u$, $\boldsymbol{z}_v$ have at least one different component.

Denote with ρ_j the vector which contains all the α_j distinct values for the input variable x_j in the training set, arranged in ascending order, i.e. $\rho_{jl}<\rho_{j,l+1}$ for each $l = 1, ..., \alpha_j-1$. Then, we can consider a set of binary values τ_{jl}, with $j = 1, ..., d$ and $l = 1, ..., \alpha_j-1$, asserting if a separation must be set for the j-th variable between its l-th and $(l+1)$-th values:

$$\tau_{jl} = \begin{cases} 1, \text{ if } \gamma_j \text{ contains } \rho_{jl} \\ 0, \qquad \text{otherwise} \end{cases}$$

Of course, the total number of possible cutoffs is given by

$$\sum_{j=1}^{d} \sum_{l=1}^{\alpha_j-1} \tau_{jl}$$

which must be minimized under the constraint that examples $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ belonging to different classes have to be separated at least by one cutoff. To this aim, let X_{juv} the set of indexes l such that ρ_{jl} lies between x_{uj} and x_{vj}:

$$X_{juv} = \left\{ l | x_{uj} < \rho_{jl} < x_{vj} | \right\}$$

Then, the discretization problem can be stated as:

$$min_\tau \sum_{j=1}^{d} \sum_{l=1}^{\alpha_j-1} \tau_{jl}$$

$$subj\ to \sum_{j=1}^{d} \sum_{l \in X_{juv}} \tau_{jl} \geq 1 \quad \text{for each } u,\ v, s.t. y_u \neq y_v \qquad (1)$$

To improve the separation ability of the resulting set of cutoffs the constraint in (1) can be reinforced by imposing that

$$\sum_{j=1}^{d} \sum_{l \in X j_{uv}} \tau_{jl} \geq s$$

for some $s \geq 1$. Intensive trials on real world datasets have shown that a good value for s is given by $s = 0.2d$; this choice has been adopted in all the analysis performed in the present paper.

Since the solution of the programming problem in (1) can require an excessive computational cost, a near-optimal greedy approach is adopted by the Attribute Driven Incremental Discretization (ADID) procedure [39]. It follows an iterative algorithm that adds iteratively the cutoff obtaining the highest value of a quality measure based on the capability of separating patterns belonging to different classes. Smart updating procedures enable ADID to efficiently attain a (sub) optimal discretization.

After the set of candidate cutoffs is produced, a subsequent phase is performed, to refine their position. This updating task significantly increases the robustness of final discretization.

Classification by ADID and LLM implemented in Rulex 2.0

A classification model was built on the expression values of the 62 probe sets constituting NB-hypo signature [1]. Model generation and performance was established by splitting the dataset into a training set, comprising 60% of the whole patients cohort, and a test set comprising the remaining 40%. To build a classifier, a Rulex 2.0 process was designed. A discretizer component that adopts the Attribute Driven Incremental Discretization (ADID) procedure [39] and a classification component that adopts a

rule generation method called Logic Learning Machine (LLM) were utilized into the process. Entropy based (EntMDL [42]), Modified Chi Square [43], ROC based (Highest Youden index ([44]), and Equal frequency (i.e. median expression for each feature) components have been executed as alternative discretization methods. To design the most accurate classifier one important parameter of the LLM component was evaluated. The parameter was the maximum error allowed on the training set. It defines the maximum percentage of examples covered by a rule with a class differing from the class of the rule. The parameter values evaluated ranged in the set 0%, 5%, 10%, 15%, 20%, 25%, and 30%. For each parameter value, a 10 times repeated 10-fold cross validation analysis was performed and the classification performances were collected. The parameters' choice that obtained the best mean classification accuracy was selected to train the final gene expression based classifier on the whole training set utilizing the aforementioned Rulex components. The *Rulex* software suite is commercialized by Impara srl[40].

Decision Tree [45], Support Vector Machines (SVM) [46], and Prediction Analysis of Microarrays (PAM) [47] were run on the same training and test sets for reference.

Performance evaluation in predicting patients' outcome

To evaluate the prediction performance of the classifiers we used the following metrics: accuracy, recall, specificity and negative predictive values (NPV), considering good outcome patients as positive instances and poor outcome patients as negative instances. Accuracy is the proportion of correctly predicted examples in the overall number of instances. Recall is the proportion of correctly predicted positive examples against all the positive instances of the dataset. Precision is the proportion of correctly classified positive examples against all the predicted positive instances. Specificity is the proportion of correctly predicted negative examples against all the negative instances of the dataset. NPV is the proportion of the correctly classified negative examples against all the predicted negative instances.

Rule quality measures

Rule generation methods constitute a subset of classification techniques that generate explicit models $g(\boldsymbol{x})$ described by a set of m rules r_k, $k = 1, ..., m$, in the **if-then** form:

$$\textbf{if} < \textit{premise} > \textbf{then} < \textit{consequence} >$$

where <*premise*> is the logical product (**and**) of m_k conditions c_{kl}, with $l = 1, ..., m_k$, on the components x_j, whereas <*consequence*> gives a class assignment $\gamma = \tilde{\gamma}$ for the output. In general, a condition c_{kl} in the premise involving an ordered variable x_j has one of the following forms $x_j > \lambda$, $x_j \leq \mu$, $\lambda < x_j \leq \mu$, being λ and μ two real values, whereas a nominal variable x_k leads to membership conditions $x_k \in \{\alpha, \delta, \sigma\}$, being α, δ, σ admissible values for the k-th component of $\boldsymbol{x}$.

For instance, if x_1 is an ordered variable in the domain {1, ..., 100} and x_2 is a nominal component assuming values in the set {*red, green, blue*}, a possible rule r_1 is

$$\textbf{if } x_1 > 40 \textbf{ and } x_2 \in \{\textit{red, blue}\} \textbf{ then } \gamma = 0$$

where 0 denotes one of the q possible assignments (classes).

According to the output value included in their consequence part, the m rules r_k describing a given model $g(\boldsymbol{x})$ can be subdivided into q groups $G_1, G_2, ..., G_q$. Considering the training set S, any rule $r \in G_l$ is characterized by four quantities: the numbers $TP(r)$ and $FP(r)$ of examples $(\boldsymbol{x}_i, y_i)$ with $y_i = y_l$ and $y_i \neq y_l$, respectively, that satisfy all the conditions in the premise of r, and the numbers $FN(r)$ and $TN(r)$ of examples $(\boldsymbol{x}_i, y_i)$ with $y_i = y_l$ and $y_i \neq y_l$, respectively, that do not satisfy at least one of the conditions in the premise of r.

The quality of a rule was measured utilizing the following quantities. Give a rule r, we define the *covering* $C(r)$, the *error* $E(r)$, and the *precision* $P(r)$ according to the following formulas:

$$C(r) = \frac{TP(r)}{TP(r) + FN(f)} \quad , \quad E(r) = \frac{FP(r)}{FP(r) + TN(r)}, \quad P(r) = \frac{TP(r)}{TP(r) + FP(r)}$$

The covering of a rule is the fraction of examples in the training set that satisfy the rule and belong to the target class. The error of a rule is the fraction of examples in the training set that satisfy the rule and do not belong to the target class. The precision of a rule is the fraction of examples in the training set that do not belong to the target class but satisfy the premises of the rule. The greater was the covering and the precision, the higher was the generality and the correctness of the corresponding rule.

To test the statistical significance of the rules we used a Fisher's exact test (FET) implemented by the software package R. The test of significance considered significant any rule having $P < 0.05$.

Relevance measure and ranking of the probe sets

To obtain a measure of importance of the features included into the rules and rank these features according to this value, we utilized a measure called Relevance $R(c)$ of a condition c. Consider the rule r' obtained by removing that condition from r. Since the premise part of r' is less stringent, we obtain that $E(r') \geq E(r)$ so that the quantity $R(c) = (E(r') - E(r))C(r)$ can be used as a measure of relevance for the condition c of interest.

Since each condition c refers to a specific component of $\boldsymbol{x}$, we define the relevance $R_v(x_j)$ for every input

variable x_j as follows:

$$R_v(x_j) = 1 - \prod_k (1 - R(c_{kl}))$$

where the product is computed on the rules r_k that includes a condition c_{kl} on the variable x_j.

Denote with V_{kl} the attribute involved in the condition c_{kl} of the rule r_k and with S_{kl} the subset of values of V_{kl} for which the condition c_{kl} is verified. If V_{kl} is an ordered attribute and the condition c_{kl} is $V_{kl} \leq a$ for some value $a \in S_{kl}$, then the contribution to $R_v(x_j)$ is negative. Hence, by adding the superscript – (resp. +) to denote the attribute V_{kl} with negative (resp. positive) contribution, we can write $R_v(x_j)$ for an ordered input variable x_j in the following way:

$$R_v(x_j) = \prod_{V_{kl}} (1 - R(c_{kl})) - \prod_{V_{kl}} (1 - R(c_{kl}))$$

where the first (resp. second) product is computed on the rules r_k that includes a condition c_{kl}

leading to a negative (resp. positive) contribution for the variable x_j.

Output assignment for a new instance

When the model g(**x**) described by the set of m rules r_k, k = 1, ..., m, is employed to classify a new instance **x**, the <premise> part of each rule is examined to verify if the components of **x** satisfy the conditions included in it. Denote with Q the subset of rules whose <premise> part is satisfied by **x**; then, the following three different situations can occur:

1. The set Q includes only rules having the same output value $\tilde{y}$ in their <consequence> part; in this case the class $\tilde{y}$ is assigned to the instance **x**.
2. The set Q contains rules having different output values in their <consequence> part; it follows that Q can be partitioned into q disjoint subsets Q_i, (some of which can be empty) including the rules r pertaining to the ith class. In this case, to every attribute x_j can be assigned a measure of consistency t_{ij} given by the maximum of the relevance r(c) for the conditions c involving the attribute x_j and included in the <premise> part of the rules in Q_i. Then, to the instance **x** is assigned the class $\tilde{y}$ associated with the following maximum:

$$\tilde{y} = \underset{i=1,\cdots q}{\operatorname{argmax}} \sum_{i=1}^{d} t_{ij}$$

3. The set Q is empty, i.e. no rule is satisfied by the instance $\boldsymbol{x}$; in this case the set Q_{-1} containing the subset of rules whose <*premise*> part is satisfied by $\boldsymbol{x}$ except for one condition is considered and points 1 and 2 are again tested with $Q = Q_{-1}$. If again Q is empty the set Q_{-2} containing the subset of rules whose <*premise*> part is satisfied by $\boldsymbol{x}$ except for two conditions is considered and so on.

The conflicting case 2 can be controlled in Rulex 2.0 by assigning a set of weights w_i to the output classes; in this way equation (1) can be written as

$$\tilde{y} = \underset{i=1,\cdots q}{\operatorname{argmax}} \sum_{i=1}^{d} t_{ij} w_i$$

and we can speak of weighted classification.

Additional material

Additional file 1: Title of data: Batch effect and LLM prediction performance. Description of data: the file contains a table showing the influence of batch effect on LLM prediction performance. Additional file 1. Table 1. *Influence of batch effect on LLM prediction performance*. The table shows the influence of batch effect calculated on accuracy, recall, precision, and specificity and NPV measures. Performances are comparable removing batch effect from the dataset.

List of abbreviations

INSS: International Neuroblastoma Staging System; FET: Fisher's Exact test; NPV: Negative Predictive Value; INRG: International Neuroblastoma Risk Group; LLM: Logic Learning Machine; SNN: Switching Neural Networks; SC: Shadow Clustering; TP: true positives; FP: false positives; TN: true negatives; FN: false negatives; NB: neuroblastoma; ADID: Attribute Driven Incremental Discretization; WCS: weighted classification system; PVCA: principal variance component analysis; SVA: surrogate variable analysis; FSVA: frozen surrogate variable analysis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

DC conceived the project, performed the statistical analysis and drafted the manuscript. MM suggested the use of LLM, designed some of the experiments, designed the Rulex software and helped to draft the manuscript. SP performed computer experiments and helped to draft the manuscript, RV and MC, participated to the development of the project. FB and PB carried out the microarray data analysis. LV supervised the study and wrote the manuscript.

Acknowledgements

The work was supported by the Fondazione Italiana per la Lotta al Neuroblastoma, the Associazione Italiana per la Ricerca sul Cancro, the Società Italiana Glicogenosi, the Fondazione Umberto Veronesi, the Ministero della Salute Italiano and the Italian Flagship Project "InterOmics". The authors would like to thank the Italian Association of Pediatric Hematology/Oncology (AIEOP) for tumor samples collection and Dr. Erika Montani for her valuable support concerning the use of both statistical and graphical Rulex 2.0 routines. DC and FB have a fellowship from the Fondazione Italiana per la Lotta al Neuroblastoma.

Declarations

Charge for this article was paid by a grant of the Fondazione Italiana per la Lotta al Neuroblastoma.

This article has been published as part of *BMC Bioinformatics* Volume 15 Supplement 5, 2014: Italian Society of Bioinformatics (BITS): Annual Meeting 2013. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/15/S5

Authors' details
[1]Laboratory of Molecular Biology, Gaslini Institute, Largo Gaslini 5, 16147 Genoa, Italy. [2]Institute of Electronics, Computer and Telecommunication Engineering, National Research Council of Italy, Genoa 16149, Italy. [3]Department of Human Genetics, Academic Medical Center, University of Amsterdam, Meibergdreef 15, Amsterdam 1100, The Netherlands. [4]Department of Hematology-Oncology, Gaslini Institute, Largo Gaslini 5, Genoa 16147, Italy.

Published: 6 May 2014

References

1. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Versteeg R, Caron HN, Molenaar JJ, Ora I, Eva A, *et al*: **A biology-driven approach identifies the hypoxia gene signature as a predictor of the outcome of neuroblastoma patients.** *Molecular Cancer* 2010, **9**:185.
2. Thiele CJ: **Neuroblastoma.** In *Human Cell Culture.* London: Kluwer Academic;Master JRW, Palsson B 1999:21-22.
3. Haupt R, Garaventa A, Gambini C, Parodi S, Cangemi G, Casale F, Viscardi E, Bianchi M, Prete A, Jenkner A, *et al*: **Improved survival of children with neuroblastoma between 1979 and 2005: a report of the Italian Neuroblastoma Registry.** *J Clin Oncol* 2010, **28**:2331-2338.
4. Doroshow JH: **Selecting systemic cancer therapy one patient at a time: is there a role for molecular profiling of individual patients with advanced solid tumors?** *J Clin Oncol* 2010, **28**:4869-4871.
5. Wei J, Greer B, Westermann F, Steinberg S, Son C, Chen Q, Whiteford C, Bilke S, Krasnoselsky A, Cenacchi N, *et al*: **Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma.** *Cancer Res* 2004, **64**:6883-6891.
6. Schramm A, Schulte JH, Klein-Hitpass L, Havers W, Sieverts H, Berwanger B, Christiansen H, Warnat P, Brors B, Eils J, *et al*: **Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling.** *Oncogene* 2005, **24**:7902-7912.
7. Ohira M, Oba S, Nakamura Y, Isogai E, Kaneko S, Nakagawa A, Hirata T, Kubo H, Goto T, Yamada S: **Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas.** *Cancer Cell* 2005, **7**:337-350.
8. Oberthuer A, Berthold F, Warnat P, Hero B, Kahlert Y, Spitz R, Ernestus K, Konig R, Haas S, Eils R, *et al*: **Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification.** *J Clin Oncol* 2006, **24**:5070-5078.
9. Fischer M, Oberthuer A, Brors B, Kahlert Y, Skowron M, Voth H, Warnat P, Ernestus K, Hero B, Berthold F: **Differential expression of neuronal genes defines subtypes of disseminated neuroblastoma with favorable and unfavorable outcome.** *Clin Cancer Res* 2006, **12**:5118-5128.
10. Vermeulen J, De Preter K, Naranjo A, Vercruysse L, Van Roy N, Hellemans J, Swerts K, Bravo S, Scaruffi P, Tonini GP, *et al*: **Predicting outcomes for children with neuroblastoma using a multigene-expression signature: a retrospective SIOPEN/COG/GPOH study.** *Lancet Oncol* 2009, **10**:663-671.
11. De Preter K, Vermeulen J, Brors B, Delattre O, Eggert A, Fischer M, Janoueix-Lerosey I, Lavarino C, Maris JM, Mora J, *et al*: **Accurate Outcome Prediction in Neuroblastoma across Independent Data Sets Using a Multigene Signature.** *Clin Cancer Res* 2010, **16**:1532-1541.
12. Oberthuer A, Hero B, Berthold F, Juraeva D, Faldum A, Kahlert Y, Asgharzadeh S, Seeger R, Scaruffi P, Tonini GP, *et al*: **Prognostic Impact of Gene Expression-Based Classification for Neuroblastoma.** *J Clin Oncol* 2010, **28**:3506-3515.
13. Cornero A, Acquaviva M, Fardin P, Versteeg R, Schramm A, Eva A, Bosco MC, Blengio F, Barzaghi S, Varesio L: **Design of a multi-signature ensemble classifier predicting neuroblastoma patients' outcome.** *BMC Bioinformatics* 2012, **13(Suppl 4)**:S13.
14. Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Varesio L: **The l1-l2 regularization framework unmasks the hypoxia signature hidden in the transcriptome of a set of heterogeneous neuroblastoma cell lines.** *BMC Genomics* 2009, **10**:474.
15. Cangelosi D, Blengio F, Versteeg R, Eggert A, Garaventa A, Gambini C, Conte M, Eva A, Muselli M, Varesio L: **Logic Learning Machine creates explicit and stable rules stratifying neuroblastoma patients.** *BMC Bioinformatics* 2013, **14(Suppl 7)**:S12.
16. Pietras A, Johnsson AS, Pahlman S: **The HIF-2alpha-driven pseudo-hypoxic phenotype in tumor aggressiveness, differentiation, and vascularization.** *Curr Top Microbiol Immunol* 2010, **345**:1-20.
17. Semenza GL: **Regulation of cancer cell metabolism by hypoxia-inducible factor 1.** *Semin Cancer Biol* 2009, **19**:12-16.
18. Carmeliet P, Dor Y, Herbert JM, Fukumura D, Brusselmans K, Dewerchin M, Neeman M, Bono F, Abramovitch R, Maxwell P, *et al*: **Role of HIF-1alpha in hypoxia-mediated apoptosis, cell proliferation and tumour angiogenesis.** *Nature* 1998, **394**:485-490.
19. Lin Q, Yun Z: **Impact of the hypoxic tumor microenvironment on the regulation of cancer stem cell characteristics.** *Cancer Biol Ther* 2010, **9**:949-956.
20. Lu X, Kang Y: **Hypoxia and hypoxia-inducible factors: master regulators of metastasis.** *Clin Cancer Res* 2010, **16**:5928-5935.
21. Chan DA, Giaccia AJ: **Hypoxia, gene expression, and metastasis.** *Cancer Metastasis Rev* 2007, **26**:333-339.
22. Harris AL: **Hypoxia–a key regulatory factor in tumour growth.** *Nat Rev Cancer* 2002, **2**:38-47.
23. Rankin EB, Giaccia AJ: **The role of hypoxia-inducible factors in tumorigenesis.** *Cell Death Differ* 2008, **15**:678-685.
24. Fardin P, Cornero A, Barla A, Mosci S, Acquaviva M, Rosasco L, Gambini C, Verri A, Varesio L: **Identification of Multiple Hypoxia Signatures in Neuroblastoma Cell Lines by l(1)-l(2) Regularization and Data Reduction.** *Journal of Biomedicine and Biotechnology* 2010.
25. Kotsiantis SB, Zaharakis ID, Pintelas PE: **Machine learning: a review of classification and combining techniques.** *Artif Intell Rev* 2006, **26**:159-190.
26. Tan AC, Naiman DQ, Xu LF, Winslow RL, Geman D: **Simple decision rules for classifying human cancers from gene expression profiles.** *Bioinformatics* 2005, **21**:3896-3904.
27. Fürnkranz J: **Separate-and-conquer rule learning.** *Artificial Intelligence Review* 1999, **13**:3-54.
28. Muselli M, Ferrari E: **Coupling Logical Analysis of Data and Shadow Clustering for Partially Defined Positive Boolean Function Reconstruction.** *IEEE Transactions on Knowledge and Data Engineering* 2011, **23**:37-50.
29. Muselli M, Liberati D: **Binary rule generation via Hamming Clustering.** *IEEE Transactions on Knowledge and Data Engineering* 2002, **14**:1258-1268.
30. Boros E, Hammer P, Ibaraki T, Kogan A, Muchnik I: **An implementation of logical analysis of data.** *IEEE Transactions on Knowledge and Data Engineering* 2000, **12**:292-306.
31. Muselli M, Costacurta M, Ruffino F: **Evaluating switching neural networks through artificial and real gene expression data.** *Artif Intell Med* 2009, **45**:163-171.
32. Mangerini R, Romano P, Facchiano A, Damonte G, Muselli M, Rocco M, Boccardo F, Profumo A: **The application of atmospheric pressure matrix-assisted laser desorption/ionization to the analysis of long-term cryopreserved serum peptidome.** *Anal Biochem* 2011, **417**:174-181.
33. Muselli M: **Switching Neural Networks: A New Connectionist Model for Classification.** In *WIRN/NAIS 2005 Volume 3931.* Berlin: Springer-Verlag; Apolloni B, Marinaro M, Nicosia G, Tagliaferri R 2006:23-30.
34. Rocco CM, Muselli M: **Approximate multi-state reliability expressions using a new machine learning technique.** *Reliability Engineering and System Safety* 2005, **89**:261-270.
35. Zambrano O, Rocco CM, Muselli M: **Estimating female labor force participation through statistical and machine learning methods: A comparison.** In *Computational Intelligence in Economics and Finance Volume 2.* Berlin: Springer- Verlag;Shu-Heng C, Paul P W, Tzu-Wen K 2007:93-106.
36. Rocco CM, Muselli M: **Machine learning models for bulk electric system well-being assessment.** *12th Conference of the Spanish Association for Artificial Intelligence* CAEPIA; 2007.
37. Paoli G, Muselli M, Bellazzi R, Corvo R, Liberati D, Foppiano F: **Hamming clustering techniques for the identification of prognostic indices in patients with advanced head and neck cancer treated with radiation therapy.** *Med Biol Eng Comput* 2000, **38**:483-486.
38. Ferro P, Forlani A, Muselli M, Pfeffer U: **Alternative splicing of the human estrogen receptor alpha primary transcript: mechanisms of exon skipping.** *Int J Mol Med* 2003, **12**:355-363.
39. Ferrari E, Muselli M: **Maximizing pattern separation in discretizing continuous features for classification purposes.** *The 2010 International Joint Conference on Neural Networks (IJCNN)* 2010, **2010**:1-8, 18 July.
40. ***Rulex* software suite.** [http://www.impara-ai.com].
41. Boedigheimer MJ, Wolfinger RD, Bass MB, Bushel PR, Chou JW, Cooper MF, *et al*: **Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories.** *BMC Genomics* 2008, **9**:285.

42. Kohavi R, Sahami M: **Error-Based and Entropy-Based Discretization of Continuous Features.** *Procedings of the Second International Conference on Knowledge Discovery and Data Mining* AAAI Press; 1996, 114-119.
43. Tay FEH, Shen L: **A Modified Chi2 Algorithm for Discretization.** *IEEE Transactions on Knowledge and Data Engineering* 2002, **14**:666-670.
44. Perkins NJ, Schisterman EF: **The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve.** *Am J Epidemiol* 2006, **163**:670-675.
45. Quinlan JR: *C4.5: programs for machine learning* Morgan Kaufmann Publishers Inc; 1993.
46. Chang C, Lin C: **LIBSVM: A library for support vector machines.** *ACM Transactions on Intelligent Systems and Technology* 2011, **2**:1-27.
47. Tibshirani RF, Hastie TF, Narasimhan BF, Chu G: **Diagnosis of multiple cancer types by shrunken centroids of gene expression.** *Proc Natl Acad Sci USA* 2002, **99**:6567-6572.
48. Blagus RF, Lusa L: **Class prediction for high-dimensional class-imbalanced data.** *BMC Bioinformatics* 2010, **11**:523.
49. Takemura AF, Shimizu AF, Hamamoto K: **A cost-sensitive extension of AdaBoost with markov random field priors for automated segmentation of breast tumors in ultrasonic images.** *Int J Comput Assist Radiol Surg* 2010, **5**:537-547.
50. Vidrighin CF, Potolea R: **ProICET: a cost-sensitive system for prostate cancer data.** *Health informatics journal* 2008, **14**:297-307.
51. Sun TF, Zhang RF, Wang JF, Li XF, Guo X: **Computer-aided diagnosis for early-stage lung cancer based on longitudinal and balanced data.** *PLoS ONE* 2013, **8**:e63559.
52. Doyle S, Monaco JF, Feldman MF, Tomaszewski JF, Madabhushi A: **An active learning based classification strategy for the minority class problem: application to histopathology annotation.** *BMC Bioinformatics* 2011, **12**:424.
53. Teramoto R: **Balanced gradient boosting from imbalanced data for clinical outcome prediction.** *Stat Appl Genet Mol Biol* 2009, **8**:1544-6115, (Electronic).
54. Maxwell P, Pugh C, Ratcliffe P: **Activation of the HIF pathway in cancer.** *Current Opinion in Genetics & Development* 2001, **11**:293-299.
55. Oberthuer A, Warnat P, Kahlert Y, Westermann F, Spitz R, Brors B, Hero B, Eils R, Schwab M, Berthold F, *et al*: **Classification of neuroblastoma patients by published gene-expression markers reveals a low sensitivity for unfavorable courses of MYCN non-amplified disease.** *Cancer Letters* 2007, **250**:250-267.
56. Schramm A, Mierswa I, Kaderali L, Morik K, Eggert A, Schulte JH: **Reanalysis of neuroblastoma expression profiling data using improved methodology and extended follow-up increases validity of outcome prediction.** *Cancer Lett* 2009, **282**:55-62.
57. Mccall MN, Almudevar A: **Affymetrix GeneChip microarray preprocessing for multivariate analyses.** *Briefings in Bioinformatics* 2012, **13**:536-546.
58. Upton GJG, Harrison AP: **Motif effects in Affymetrix GeneChips seriously affect probe intensities.** *Nucleic Acids Res* 2012, **40**:9705-9716.
59. Brown JM, William WR: **Exploiting tumour hypoxia in cancer treatment.** *Nat Rev Cancer* 2004, **4**:437-447.
60. Johnson WE, Li CF, Rabinovic A: **Adjusting batch effects in microarray expression data using empirical Bayes methods.** *Biostatistics* 2007, **8**:118-127.
61. Luo JF, Schumacher MF, Scherer AF, Sanoudou DF, Megherbi DF, Davison TF, Shi TF, Tong WF, Shi LF, Hong HF, *et al*: **A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data.** *Pharmacogenomics J* 2010, **10**:278-291.
62. Huang S, Laoukili J, Epping MT, Koster J, Holzel M, Westerman BA, Nijkamp W, Hata A, Asgharzadeh S, Seeger RC, *et al*: **ZNF423 is critically required for retinoic acid-induced differentiation and is a marker of neuroblastoma outcome.** *Cancer Cell* 2009, **15**:328-340.
63. Ohtaki M, Otani K, Hiyama K, Kamei N, Satoh K, Hiyama E: **A robust method for estimating gene expression states using Affymetrix microarray probe level data.** *BMC Bioinformatics* 2010, **11**:183.
64. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD: **The sva package for removing batch effects and other unwanted variation in high-throughput experiments.** *Bioinformatics* 2012, **28**:882-883.

doi:10.1186/1471-2105-15-S5-S4
Cite this article as: Cangelosi *et al.*: Use of Attribute Driven Incremental Discretization and Logic Learning Machine to build a prognostic classifier for neuroblastoma patients. *BMC Bioinformatics* 2014 **15**(Suppl 5):S4.