Aggressive and Effective Feature Selection using Genetic Programming

Isac Sandin, Guilherme Andrade, Felipe Viegas,
Daniel Madeira and Leonardo Rocha
Department of Computer Science
Universidade Federal de São João del Rei
São João del Rei, Minas Gerais, Brazil
{isacsandin,gandrade,fviegas,dmadeira,lcrocha}@ufsj.edu.br

Thiago Salles and Marcos Gonçalves
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Minas Gerais, Brazil
{tsalles,mgoncalv}@dcc.ufmg.br

Abstract—One of the major challenges in automatic classification is to deal with highly dimensional data. Several dimensionality reduction strategies, including popular feature selection metrics such as Information Gain and χ^2 , have already been proposed to deal with this situation. However, these strategies are not well suited when the data is very skewed, a common situation in real-world data sets. This occurs when the number of samples in one class is much larger than the others, causing common feature selection metrics to be biased towards the features observed in the largest class. In this paper, we propose the use of Genetic Programming (GP) to implement an aggressive, yet very effective, selection of attributes. Our GP-based strategy is able to largely reduce dimensionality, while dealing effectively with skewed data. To this end, we exploit some of the most common feature selection metrics and, with GP, combine their results into new sets of features, obtaining a better unbiased estimate for the discriminative power of each feature. Our proposal was evaluated against each individual feature selection metric used in our GP-based solution (namely, Information Gain, χ^2 , Odds-Ratio, Correlation Coefficient) using a k8 cancer-rescue mutants data set, a very unbalanced collection referring to examples of p53 protein. For this data set, our solution not only increases the efficiency of the learning algorithms, with an aggressive reduction of the input space, but also significantly increases its accuracy.

I. INTRODUCTION

Automatic classification is a supervised learning strategy that, given a set of previously labeled examples, learns a model for classifying new unseen examples (test set). More formally, given a training set $\mathbb{D}_{train} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, in which x_i represents an example and y_i its class, the objective is to learn a classification model that predicts the label of $\mathbb{D}_{test} = \{(x'_1, ?), \dots, (x'_k, ?)\}$. There exists a wide range of proposed algorithms for automatic classification, although several challenges continue to receive a significant attention from the research community, such as dealing with high dimensionality and skewed data.

Particularly, learning with data with high dimensionality (also known as the $p \gg N$ problem—where p denotes the input space dimension and N denotes the number of training examples) is undoubtedly one of the major challenges in machine learning research. As the dimensionality p of the input space increases, the number of labeled instances required to produce an accurate enough model also increases, but in an exponential fashion [1]. This requirement becomes a

critical factor which potentially limits the applicability of learning techniques in real world problems. In essence, when learning with data with high dimensionality the identification of patterns and/or dense regions in the input space may become a very complex task, motivating the use of dimensionality reduction techniques prior to the use of machine learning algorithms. Several dimensionality reduction strategies have already been proposed, such as the commonly used feature selection methods. In fact, these methods are essential to guarantee efficient and accurate learning tasks. A number of feature selection metrics to assess the feature's discriminative power (and ultimately guide the selection process) have been explored, such as Information Gain [2], [3], χ^2 [4], Odds-Ratio [5], Correlation Coefficient [2], [6], Bi-Normal Separation [4], among others. These metrics estimate a score for each feature (i.e., dimensions of the input space) in order to measure its importance in pattern discrimination [4]. Hence, an effective selection of features contributes not only to the learning efficiency, due to the lower memory and processing demands, but also to the learning accuracy, since less informative or noise features are filtered out, decreasing the required number of examples to ensure an accurate model.

In addition to the high dimensionality problem, real world data sets are often skewed. This problem has been receiving an increasing attention by the data mining community [7] due to its wide spread in real world problems. The skewed data problem refers to the case where the number of examples belonging to one class largely outnumbers the other classes, causing some learning algorithms to be biased towards the larger class. In several problems, however, it is more important to discriminate examples from the minority classes, highlighting the challenge of dealing with such skewed situations. Besides hampering learning effectiveness, skewed data also makes feature selection harder. It has already been shown [8] that metrics commonly used for this purpose are biased and may not work properly with skewed data. This makes learning tasks even more challenging. As we shall discuss in Section II, to the best of our knowledge there is no agreement towards a proper use of feature selection solutions in this scenario, motivating us to combine widely used metrics in order to come up with a better selection in face of skewed data. As

validated in [4], “feature selection should then be relatively more important in difficult, high-skew situations”.

In this paper we propose a general solution for a more effective feature selection strategy, which in addition to providing a highly effective selection of important features, is also robust when dealing with skewed data. For this, we propose the use of Genetic Programming (GP) to learn a “compound” feature selection metric, able to take advantage of several basic metrics, in order to provide a better selection of features in face of skewed data. Our solution proposes to seek throughout the space of possible combinations of a fixed set of basic metrics (e.g., Information Gain, χ^2 , Odds Ratio) to determine an unbiased estimator of the discriminative power of the feature.

This paper is organized as follows: Section II discusses some related work and briefly describes the explored feature selection metrics that form the basis of our solution. In Section III we describe our proposed GP-based strategy to determine an unbiased feature selection metric better suited to handle skewed data sets. Our experimental evaluation is then reported in Section IV. Finally, we conclude and discuss some future work in Section V.

II. RELATED WORK

In this section we briefly review work aimed at analyzing feature selection metrics, mainly in face of skewed data. A feature selection metric is used to assign a score to each feature in order to assess its importance in the learning task. These scores are usually a function of the probabilities summarized in Table I. In the following, we refer to the features with higher probability of being observed in examples of positive classes as positive features, and those features with higher probability of not being observed in examples of positive classes as negative features.

TABLE I
PROBABILITIES USED BY TRADITIONAL FEATURE SELECTION METRICS.

<i>Positive</i> <i>Class = c_i</i>		Feature	
		f_j	Not f_j
Class	c_i	$P(c_i, f_j)$	$P(c_i, \bar{f}_j)$
	Not c_i	$P(\bar{c}_i, f_j)$	$P(\bar{c}_i, \bar{f}_j)$

An extensive study regarding twelve feature selection metrics can be found in [4]. In that work, the effectiveness of each metric was contrasted considering a benchmark of 229 text classification problem instances. Furthermore, the effects of data skewness were also studied and it was shown that standard metrics, such as Information Gain, were adversely impacted by it. The authors also proposed an alternative metric, called Bi-Normal Separation which performed consistently better than traditional metrics, even in face of skewed data.

In [8], the authors analyzed certain biases associated with the metrics Information Gain, χ^2 , Odds-Ratio and Correlation Coefficient. According to the authors, feature selection metrics can be grouped into one-sided and two-sided metrics. The first set corresponds to metrics which select the features

most indicative of positive class membership (i.e., positive features), while the second corresponds to the metrics which select features either indicative of negative class membership or positive class non-membership (i.e., negative features). It has become well accepted that negative features are indeed important to learn accurate models, since they can contribute, for example, to a higher recall [4], thus limiting the applicability of one-sided metrics, which may filter them out entirely. Regarding the two-sided metrics, when dealing with balanced data sets, both positive and negative features are selected with a proportion similar to the observed distribution of positive and negative features. However, when dealing with skewed data that does not hold: common two-sided metrics become biased towards the positive features. Indeed, in [8] it was shown that the two-sided metrics Information Gain and χ^2 were biased towards the positive features, which may not be the most effective selection. In fact, the selected features belong, with higher probability, to the larger classes (since they typically have more features than the smaller ones), which makes the statistical values of the metrics to be biased towards the larger classes. In order to address this issue, the authors proposed a wrapper model to explicitly combine positive and negative features selected by a base metric, finding an optimal ratio between them. It was shown experimentally that such wrapper model is able to produce a more effective feature selection. However, no pattern regarding the optimal ratio was found neither recommended, since it depends not only of data set but also of the learning algorithm and the feature selection metric used as a base metric. Clearly, finding such optimal proportion between positive and negative features is not an easy task.

In [9] the authors proposed a feature selection metric which aims at achieving strong performance on skewed data, by evaluating features by the Area Under the ROC Curve (AUC). More specifically, the proposed technique learns a linear classifier considering each single feature, evaluates the classifier at distinct boundary regions (that is, varying the classification threshold) and computes the AUC. Each feature is then ranked according to this metric. The authors came up with an unbiased estimator to assess the discriminative power of features which is better suited to handle skewed data. Following the direction of providing unbiased estimators for feature scoring, in [10] the authors proposed some modifications over the previously defined Gini-Index Text (GIT) [11] metric for scoring features in text classification. The authors studied a series of drawbacks of this metric when exposed to skewed data and outlined some modifications in order to come up with an unbiased estimator, which they called Complete GIT feature selection. This proposal, according to the reported experimental results, leads to a more effective selection of features, being robust to skewed data.

Motivated by the challenge of determining an unbiased estimator for feature scoring, which works well even in highly skewed settings, we aim at defining a “compound” feature selection metric, which takes into account the characteristics of a variety of widely used basic feature selection metrics. Our

Fig. 1. Example individual: $(+ (* 3 4) 5) = (3 * 4) + 5$.

proposal is to employ Genetic Programming in order to find a way to combine such basic metrics efficiently. In the next section, we detail the strategy.

III. PROPOSED SOLUTION

In this section we outline our proposal to model the problem of feature selection using Genetic Programming (GP). First, we motivate the use of such evolutionary approach to the problem of feature selection. Then, we describe in general terms the operation of a GP algorithm. Finally, we detail our modeling strategy to tackle the target problem of feature selection.

As highlighted in previous sections, there are many feature selection metrics proposed in literature, such as Information Gain, χ^2 , Odds-Ratio, Correlation Coefficient and Bi-Normal Separation. Each of these metrics may select different sets of features, since they exploit different criteria in their selection process. We emphasize that these different sets selected by these metrics may contain good discriminative features as well as not so relevant ones. Furthermore, good features selected by one method will not necessarily be selected by a second method. Thus, our hope is that, by combining the basic metrics, we might find an unbiased estimator which achieves a better (near-optimal) proportion of selected positive and negative features, retaining the most informative ones. Clearly, the search space over all possible combinations of basic feature selection metrics is very large, and a brute force search is not a wise choice. So, our proposal is to employ GP to guide the search more efficiently.

GP is an evolutionary algorithm extensively (and successfully) used to solve complex optimization and learning problems, where the search space for an optimal solution is prohibitively large. It works by mimicking the evolution process of a population of individuals, following the Darwin's principle of *survival of the fittest*. In this framework, each individual is usually represented by a tree, consisting of terminals (leaf nodes) and non-terminals (functions, e.g., arithmetic or logic operators). In Fig. 1 we illustrate an arithmetic operation represented by a tree.

The evolution of the population is driven by the generation and the combination of its individuals, according to their fitness. The following main steps summarizes the evolution process of a population in a GP framework:

- 1) Random generation of individuals (initial population), using the functions and terminals available. A widely

used strategy is the so called ramped half-and-half generation [12].

- 2) Iterative generation of a new population through the following sub-steps, until a stopping criterion is met:
 - a) Assessment of each individuals fitness, according to the problem at hand (i.e., the quality of the solution represented by them).
 - b) Probabilistic selection (based on fitness) of one or two individuals from the population to participate in the genetic operations detailed in (c). A commonly used strategy is the tournament selection, which will be employed here.
 - c) Creation of a new individual using any of the following genetic operators:
 - i) Reproduction: The selected individuals are copied to the new population.
 - ii) Mutation: A new individual is created and added to the new population after a random change in some node of the tree.
 - iii) Crossover: A new individual is created and added to the new population after recombining parts of two randomly selected individuals (trees).
- 3) In each population, the best individual created is saved as the result of the iteration. Once satisfied the stopping criterion, the best program produced is designated as the (perhaps approximate) solution to the problem.

Recall that the problem we aim to solve is to determine subsets of features which better capture the characteristics of the classes, ultimately yielding a more compact representation of the input examples (i.e., a reduced dimensionality) without decreasing (potentially, even increasing) the effectiveness of learning algorithms when discriminating between classes. To adapt the GP framework to this problem, we model each individual as a possible combination between a set of basic feature selection metrics, and let the GP search for individuals which yield the most effective combination of metrics.

More specifically, consider the individual tree representation adopted by the GP framework. Each terminal node consists of a (basic) feature selection metric, which returns a set of features considered highly discriminative by that metric (i.e., a function $f : \mathbb{D} \mapsto \mathbb{S}$, where the input $\mathbb{D}$ is the training set and the output $\mathbb{S}$ is the set of most discriminative features, according to f). Two feature selection metrics f_i and f_j (sibling nodes in the tree) are combined according to a set operation specified by their parent node (non-terminal node). The set operations may be union ($\cup$), intersection ($\cap$), set difference ($\setminus$), and so on. Fig. 2 illustrates an hypothetical individual considering our modeling strategy. As one can observe, four subsets of features, selected by four feature selection metrics (f_1 , f_2 , f_3 and f_4) are combined, generating a final subset of features, which we hope will better contribute not only to the learning efficiency, but also to its effectiveness.

During the evolution process of GP, the quality of every individual is evaluated according to a pre-defined (problem

Fig. 2. Hypothetical individual under our modeling strategy: $(\cap(\cup f_1 f_2) \setminus f_3 f_4)) = (f_1 \cup f_2) \cap (f_3 \setminus f_4)$.

dependent) fitness function. Recall that our main goal is to come up with a subset of features that, besides aggressively reducing the input dimensionality, yields a better learning effectiveness. Hence, we define the fitness function according to the classification quality, evaluated after filtering the input examples according to the subset of features selected by the individual (e.g., obtained after applying all operators of the individual). The GP framework, thus, tries to maximize the fitness function, generating individuals such that the associated subset of features lead to an improved classification. In the following, we present the experimental setup (that is, the explored data set, the set of basic feature selection metrics and the adopted classifier) as well as report the evaluation of our approach.

IV. EXPERIMENTAL EVALUATION

In this section, we outline the experimental setup adopted to evaluate our approach. We start by discussing some of the configuration details of the experiments, such as the libraries used, the basic feature selection metrics explored, and the metrics used to assess the classification quality achieved when applying those feature selection metrics (as well as the metric generated by the GP framework, as previously detailed). Next, we report and analyze the experimental results obtained.

A. Experimental setup

1) *Data set*: We adopt in our experiment the k8 cancer-rescue mutants data set [13], a highly skewed data set with examples of *p53* proteins, characterized by 5,408 features composed by: 2D electrostatic and surface based features (4,826 attributes) and 3D distance based features (582 attributes). This collection is composed by 16,715 samples, classified as active (143 samples) and inactive (16,572 samples), where “active” and “inactive” denotes the functional state of the *p53* mutants. Note that, when analyzing this data set, the ultimate goal (in a classification setting) is to discriminate the minority (i.e., “active”) samples.

2) *Feature Selection Metrics*: The following feature selection metrics were used: Information Gain [2], [3], χ^2 [4], Odds-Ratio [5], Correlation Coefficient [2], [6] and Bi-Normal Separation [4].

The following notational convention will be used hereafter: we denote t and $\bar{t}$ as the presence or absence of an attribute, respectively; the positive (“active”) and negative (“inactive”) classes are denoted by c and $\bar{c}$. The probability of a feature t to

occur in a class c is represented by $P(t|c)$. The marginal probabilities $P(c)$ and $P(t)$ indicate the probability of observing a class and a feature, respectively. All these probabilities are estimated from the training set, through Maximum Likelihood Estimates (MLE). Finally, N denotes the number of input examples. A brief description of each adopted feature selection metric is given below.

- **Information Gain (IG)**: quantifies how much information we obtain about a class when we know that a certain attribute exists or not in a sample. It is a two-sided metric defined as follows:

$$IG(t, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t' \in \{t, \bar{t}\}} P(t', c) \log \frac{P(t', c)}{P(t') \cdot P(c)} \quad (1)$$

- **Chi-square χ^2** : used in statistical analysis to test whether two events are independent. In the context of feature selection, it is used to measure the association between attributes and classes. It is two-sided metric defined as:

$$\chi^2(t, c_i) = \frac{N[P(t, c_i)P(\bar{t}, \bar{c}_i) - P(t, \bar{c}_i)P(\bar{t}, c_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)} \quad (2)$$

- **Correlation Coefficient (CC)**: used to estimate the correlation between classes and the interrelation among features. It is a variation of the Chi-square metric, where $CC = \sqrt{\chi^2}$. This metric can be seen as a one-sided Chi-square metric, defined as:

$$CC(t, c_i) = \frac{\sqrt{N}[P(t, c_i)P(\bar{t}, \bar{c}_i) - P(t, \bar{c}_i)P(\bar{t}, c_i)]}{\sqrt{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)}} \quad (3)$$

- **Odds Ratio (OR)** measures the ratio of the odds of a feature to occur in the positive class to that of the negative class. It is a one-sided metric defined as:

$$OR(t, c_i) = \log \frac{P(t|c_i)[1 - P(t|\bar{c}_i)]}{[1 - P(t|c_i)]P(t|\bar{c}_i)} \quad (4)$$

In our GP modeling of the problem, we consider the terminal nodes as feature sets selected by IG, χ^2 , CC and OR metrics. These metrics were chosen because they are widely used in the literature and show distinct characteristics. The fact that they are both one-sided metrics (CC and OR) and two-sided (IG and χ^2) was also determinant for this choice: this makes the search for the best set of features to be performed by the GP consider both the positive and negative features.

3) *Classification Algorithm and Evaluation Metrics*: As we are dealing with continuous attributes, the automatic classification algorithm used in our experiments was the Gaussian Naïve Bayes [1]. This classifier was selected due to its high efficiency, since, during the GP iterations, several classification tasks are executed and evaluated. We note here that the proposed method is general enough to be instantiated with any other classifier, such as the SVM. In order to assess the effectiveness of our approach, we use standard measures

TABLE II
GP PARAMETERS

Parameter	Value
Population Size	100
Number of Generations	30
Creation Type	Ramped Half-and-Half
Crossover Probability	98
Selection Type	Tournament
Tournament Size	7
Swap Mutation Probability	50
Shrink Mutation Probability	50
Elitism	True

adopted by the information retrieval and data mining community, namely, MicroF_1 and MacroF_1 . MicroF_1 measures the global effectiveness in terms of all decisions made by the classifier (that is, the inverse of error rate). The MacroF_1 , on the other hand, measures the classification effectiveness regarding each class independently, by computing the F_1 measure (i.e., harmonic mean between precision and recall) obtained for each class and averaging them [14].

4) *Fitness Function*: As mentioned in Section III, the fitness function is problem-dependent and must reflect the ultimate goal of the optimization/learning problem. Here, our main goal is to provide a high quality classification. For this purpose, our fitness function must be a metric to assess the classification quality, enabling the GP framework to search for individuals that maximize it. As our goal in this paper is to deal with highly skewed data, we used the MacroF_1 metric to compose the fitness function. In fact, the MacroF_1 metric better captures the classification effectiveness for each classes individually (unlike the MicroF_1 metric, which assess a global classification effectiveness). This becomes important when dealing with skewed data, since the effectiveness in discriminating the minority classes is also taken into account.

5) *Genetic Programming Library*: We used the *gpc++ v0.5.2* library [15], an efficient GP library, to implement our approach. As a generic library, only the implementation of structures closely related to the problem were needed, such as terminal nodes (in our case, the feature selection metrics), function nodes (in our case, the set operators to be applied to the set of features selected by the two children nodes) and the fitness function (in our case, a single metric to assess classification effectiveness, namely, the MacroF_1 metric).

In order to find the parameters used in our experiments, we conducted a pilot study, inspired by results reported by the author of the library [15]. We note that this study does not fit all cases (since it is problem and data dependent) but, as argued by the author, it is a good start towards an ideal setting.

First, we sampled 10% of the data set in a random fashion, keeping the original distribution of the classes. We applied each feature selection metrics considering this sampled data in order to find the discriminative power of each feature. Using the top 5% of the most discriminative features, for each

metric in isolation, we applied the GP procedure, varying the parameters in order to find the values that maximize the fitness of the individuals (that is, those that provide highest quality classification). We vary each GP parameter according to the following strategy: The population size was set from 50 up to 100, with steps of 10. The number of generations ranged from 20 to 50, with steps of 10. The crossover probability was varied from 90% to 100%, with steps of 1%, while both the swap mutation and shrink mutation probabilities were varied from 20% to 100%, with steps of 10%. Finally, the tournament size used for individual selection was chosen between 5 and 10, with steps of 1. Such parameter tuning was based on a simple experimental design where, while one of them is varied, the remaining are kept fixed. The configuration that yielded the best results is found reported in Table II, and these were the values used in our experiments. Despite the good results obtained, presented in the following section, we believe that a fine tuning of these parameters for our problem can lead to even better results. We leave this study as future work.

B. Results and Discussion

In this section we describe the conducted experiments to assess the effectiveness of our proposed method for feature selection based on GP. Then, we discuss the obtained results. The results are based on the quality of the classification of *p53* samples, obtained after filtering the features for each explored metric. We contrast them with those selected by our approach. Figure 3 illustrate the adopted process, which will be detailed next.

Fig. 3. Experimental Architecture in Detail.

First, we partition the data set into two equal parts, which we call the Test and Train partitions. The first partition corresponds to the set of examples used in the evaluation of the techniques, whereas the second partition corresponds to the set of examples used by the tuning process of the feature

selection metrics, as well as by our approach. This partitioning step was carried out in a random fashion, keeping the original distribution of the classes in each half.

In a second step, for each individual metric we find the subset of attributes with the highest discriminative power, through cross-validation over the Train set. Each of these selection metrics, when applied to a data set, score each attribute according to their discriminative power. The goal is to find the top $L\%$ features with most discriminative power (that is, highest scores) which lead to best classification effectiveness. More specifically, considering the Train set, we apply a process of hold-out 70/30 (i.e., 70% of the Train set is used for actual training and the remaining 30% is retained for the Validation set) with 10 repetitions, varying L from 5% to 100%, with steps of 5%. The best results were obtained when L was set to 10%, 15%, 10% and 10% for the metrics OR, IG, χ^2 and CC, respectively.

Having determined the L values for each metric, the third step is to find the best individuals considering the GP approach. To do so we generate the feature sets corresponding to each metric using the determined L values, which will then be used in the evolution process of the GP. Again, we split the Train set into two subsets: actual Train and Validation, according to a hold-out 70/30, with 70% and 30% of the samples, respectively. As before, we performed 10 repetitions. For each repetition, the GP evolutionary process was applied and the best individual was saved. Then, the best individuals found in all repetitions were evaluated considering the Validation set, in order to select the best one (i.e., the one with better generalization capability). This step was repeated five times, resulting in 5 best individuals. In Figure 4 we show the evolution of the average population fitness during the GP evolutionary process. Due to space constraints, we only report this analysis for 2 out of 5 individuals, and point out that similar behavior was observed for all individuals. As we can observe, as the evolution process goes on, the population quality (in terms of average fitness of the population) increases, indicating that the GP is, indeed, able to follow a good path in the search space towards a maximum (which may be local or global). As we shall see next, even if stuck in a local maximum, the achieved results in terms of classification effectiveness is better than when using the traditional metrics. The five best individuals found are reported in Table III.

TABLE III
THE FIVE BEST INDIVIDUALS FOUND.

Individual	Expression
1	$\chi^2 \setminus CC$
2	$OR \setminus IG$
3	$OR \setminus IG$
4	$CC \setminus IG$
5	$CC \setminus \chi^2$

The last step is to finally assess, considering the independent Test set, the classification effectiveness after filtering the

(a) Individual 1.

(b) Individual 2.

Fig. 4. Evaluation of the Average Population Fitness.

top $L\%$ most discriminative features found for each feature selection metric, as well as the five individuals found in the third step. To do such evaluation, we randomly selected 70% of the Test set, to perform the evaluation and 70% of the Train set to perform the training stage. This process was repeated 10 times. The MacroF₁ and MicroF₁ scores, with their corresponding standard deviation values, are reported in Tables IV and V, respectively.

As we can see on Table V, our GP-based feature selection strategy provided substantial improvements over the traditional feature selection metrics, as shown in Table IV. Each feature selection metric has advantages, but are also biased towards features belonging to the majority class, leading to potentially sub-optimal subset of features. Our GP approach, on the other hand, is capable of selecting a better set of discriminative

TABLE IV
NB RESULTS COMBINED WITH EACH FEATURE SELECTION METRIC

	L(%)	MacroF ₁ (%)	MicroF ₁ (%)
NB	100	35.66 ± 0.58	52.20 ± 1.56
NB+OR	10	49.66 ± 0.78	83.92 ± 1.16
NB+CC	10	49.00 ± 0.63	82.52 ± 0.82
NB+IG	15	44.99 ± 0.87	73.58 ± 1.68
NB+ χ^2	10	39.79 ± 1.12	61.32 ± 2.25

TABLE V
THE FIVE BEST INDIVIDUALS FOUND AND CORRESPONDING RESULTS.

	# features	MacroF ₁ (%)	MicroF ₁ (%)
Ind1	9	66.65 ± 1.65	98.59 ± 0.12
Ind2	6	60.43 ± 0.90	97.59 ± 0.09
Ind3	6	56.24 ± 0.70	97.83 ± 0.31
Ind4	5	58.37 ± 0.93	98.01 ± 0.14
Ind5	9	55.49 ± 0.75	97.82 ± 0.12

features (with a better ratio between positive and negative features, according to the observed data), avoiding the potential bias when facing skewed data. Hence, it can be said that we successfully selected a more effective set of features.

Another interesting point is the aggressive reduction on the number of features in each set, without compromising the classification effectiveness. Comparing the best individual feature selection with our GP-based approach, we reduced the features from 504 ($L = 10\%$) to only 9 features. Even when selecting less features, we improved the MacroF₁ metric from 49.66% to 66.65% and the MicroF₁ from 83.92% to 98.59%, which is as very tangible improvement (gains of 34% in MacroF₁ and 17% in MicroF₁). We can observe that our strategy can do a better job in selecting the most important features, making the classification more effective.

TABLE VI
A COMPARISON BETWEEN THE FEATURE SELECTION METRICS
CONSIDERING THE TOP-9 FEATURES.

	MacroF ₁ (%)	StdDev	MicroF ₁ (%)	StdDev
GP	66.65	1.65	98.59	0.12
OR	50.87	1.64	99.13	0.06
CC	50.15	0.72	99.14	0.05
χ^2	50.06	0.52	98.82	0.22
IG	49.83	0.40	98.37	0.55

Finally, we also compared the explored traditional feature selection metrics to our approach, considering the *same* number of most discriminative features, as reported in Table VI. Considering the 9 top ranked features by the OR metric, the learned model was able to achieve MacroF₁ and Micro F₁ of 50.87 ± 1.64 and 99.13 ± 0.06 , respectively. Contrasting these values with the ones reported in Table IV we can observe an improvement in MicroF₁, but a statistically tied MacroF₁

value. This means that, after applying the OR filter, the learned model became more biased towards the larger class. On the other hand, with the same number of features, the GP-based approach was able to achieve a significantly higher MacroF₁ value, indicating that the learned classification model, after filtering the features, was more effective to classify test examples belonging to the minority class. In that table, we can also observe that the GP approach outperformed all traditional feature selection metrics in terms of MacroF₁, indicating a more effective discrimination of the minority class examples. In fact, the GP was able to produce a much more effective set of features. This corroborates our argument that our GP based solution does a better job when selecting the best features.

V. CONCLUSIONS AND FUTURE WORK

Automatically classifying data with high dimensionality is a challenge. This challenge is even more difficult when dealing with skewed datasets since methods commonly used for reducing dimensionality by feature selection are usually biased towards the larger class, when the goal is usually to maximize classification effectiveness in the smaller classes.

In this paper we propose a general solution for a more effective feature selection strategy, which in addition to providing a highly effective selection of important features, is also robust to skewed data. For this, we propose the use of Genetic Programming (GP) to learn a “compound” feature selection metric, able to take advantage of several basic metrics. More specifically, each terminal node consists of a “basic” feature selection metric, which returns a set of features considered highly discriminative. Two feature selection metrics are combined according to a set operation specified by their parent node. Moreover, we define the fitness function according to the classification quality, evaluated after filtering the input examples according to the subset of features selected by the individual. Our experimental results show that, when applied to the *k8* cancer-rescue mutants dataset, a very unbalanced collection referring to examples of *p53* protein classified as cancer suppressor or not, our solution not only increases the efficiency of the learning algorithms (with a reduction of around 98.2% in the size of the data space, from 504 features to only 9), but also significantly increases its accuracy (with gains of 34% in MacroF₁ and 17% in MicroF₁ when compared to the best baseline) when classifying *p53* samples.

As a future work we intend to apply our solution to other data sets where the problem of high dimensionality and skewed distributions are even more significant (e.g., image and textual datasets), perform a more complete study of the GP-parameters, and compare our approach to an unbiased metric such as Bi-Normal Separation extended for continuous features.

ACKNOWLEDGMENTS

This work was partially supported by CNPq, CAPES, Fapemig, FINEP, and InWeb—Brazilian National Institute of Science and Technology for the Web.

REFERENCES

- [1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.
- [2] F. Sebastiani and C. N. D. Ricerche, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, pp. 1–47, 2002.
- [3] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the Fourteenth International Conference on Machine Learning*, ser. ICML '97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 412–420. [Online]. Available: <http://dl.acm.org/citation.cfm?id=645526.657137>
- [4] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *The Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.
- [5] D. Mladenic, "Machine learning on non-homogeneous, distributed text data." Ph.D. dissertation, University of Ljubljana, Faculty of Computer and Information Science, 1998.
- [6] H. T. Ng, W. B. Goh, and K. L. Low, "Feature selection, perceptron learning, and a usability case study for text categorization," in *Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '97. New York, NY, USA: ACM, 1997, pp. 67–73. [Online]. Available: <http://doi.acm.org/10.1145/258525.258537>
- [7] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (The Morgan Kaufmann Series in Data Management Systems)*, 1st ed. Morgan Kaufmann, Oct. 1999.
- [8] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM SIGKDD Explorations Newsletter*, vol. 6, pp. 80–89, 2004.
- [9] X.-w. Chen and M. Wasikowski, "Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08, 2008, pp. 124–132.
- [10] H. Park, S. Kwon, and H.-C. Kwon, "Complete gini-index text (git) feature-selection algorithm for text classification," in *Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on*, 2010, pp. 366–371.
- [11] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," vol. 33, no. 1, 2007.
- [12] J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems)*. Cambridge, MA, USA: The MIT Press, 1992.
- [13] S. A. Danziger, R. Baronio, L. Ho, L. Hall, K. Salmon, G. W. Hatfield, P. Kaiser, and R. H. Lathrop, "Predicting positive p53 cancer rescue regions using most informative positive (mip) active learning," *PLoS Comput Biol*, vol. 5, no. 9, 09 2009.
- [14] D. D. Lewis, "Evaluating and optimizing autonomous text classification systems," in *Eighteenth Annual, International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 1995, pp. 264–254.
- [15] T. Weinbrenner, "Genetic programming techniques applied to measurement data," Diploma Thesis, Germany, February 1997. [Online]. Available: <http://citeseer.ist.psu.edu/weinbrenner97genetic.html>