Two Simple and Effective Feature Selection Methods for Continuous Attributes with Discrete Multi-class

Manuel Mejía-Lavalle, Eduardo F. Morales, and Gustavo Arroyo

Instituto de Investigaciones Eléctricas, Reforma 113, 62490 Cuernavaca, Morelos, México
INAOE, L.E.Erro 1, 72840 StMa. Tonantzintla, Puebla, México
mlavalle@iie.org.mx, emorales@inaoep.mx, garroyo@iie.org.mx

Abstract. We present two feature selection methods, inspired in the Shannon's entropy and the Information Gain measures, that are easy to implement. These methods apply when we have a database with continuous attributes and discrete multi- class. The first method applies when attributes are independent among them given the class. The second method is useful when we suspect that interdependencies among the attributes exist. In the experiments that we realized, with synthetic and real databases, the proposed methods are shown to be fast and to produce near optimum solutions, with a good feature reduction ratio.

1 Introduction

Feature selection has shown to be promising pre-processing step in data mining because it can eliminate the irrelevant or redundant attributes that cause the mining tools to become inefficient and ineffective [1]; at the same time, it can preserve, and in many cases, increase the classification quality of the mining algorithm and help in the understanding of the induced models, as they tend to be smaller.

Although there are many feature selection algorithms reported in the specialized literature [1], none of them is perfect: some of them are effective, but very costly in computational time (e.g., *wrappers* methods [2]), and others are fast, but less effective in the feature selection task (e.g., *filter* methods [3]). Most of them need pure discrete data (e.g., nominal attributes with a nominal class) or pure continuous data (e.g., continuous attributes with a continuous class). So, if data has continuous attributes, they need to be discretized (either all the attributes or the class), however the results vary depending on the discretization method that is utilized [4].

In this article we propose two easy to implement feature selection methods that apply over continuous data with discrete class in a supervised learning context. The first method assumes that the attributes are independent among them given the class. The second method is useful when we suspect that interdependencies among the attributes exist. Both methods are inspired in the Shannon's n-dimensional entropy and the Information Gain measures. We show that the proposed methods are fast and produce near optimum solutions, selecting few attributes, according to the experiments that we realized, with synthetic and real databases.

To cover these topics, the article is organized as follows: Section 2 introduces ours feature selection methods; Section 3 details the experiments; in Section 4 we discuss

A. Gelbukh and A.F. Kuri Morales (Eds.): MICAI 2007, LNAI 4827, pp. 452–461, 2007.

and survey some works related with ours methods; conclusions and future research directions are offered in Section 5.

2 Proposed Feature Selection Methods

2.1 *vG* Method: When Attributes Are Independent

vG is inspired in Shannon's entropy and Information Gain, which emerged from the Information Theory arena. So, we begin our description introducing these basic concepts [5]. Formally, the entropy H_n of a nominal, or discrete, set of probabilities $p_1, \ldots, p_n$ has been defined as:

$$H_n = -\Sigma\ p_i\ \log_2\ p_i \tag{1}$$

Additionally, there is a less used, and known, entropy version H_c for continuous data; according to [5] is defined as:

$$H_c = -\int p(x) \log_2\ p(x)\ dx \tag{2}$$

but generally, the density distribution function $p(x)$, is unknown: Miller[6] tried to estimate this function, using Voronoi regions (a kind of discretization), but he concluded that this process has exponential complexity). As Shannon [5], physicists and statisticians point out [6], a reasonable approach is to assume that this density distribution is Gaussian, whose standard deviation is S. If we realize some algebraic manipulations over (2), assuming $p(x)$ to be Gaussian, we obtain:

$$H_c(x) = \log_2\ \{\ 2\pi e\ \}^{1/2}\ S \tag{3}$$

Observing (3) we can say that, in this terms, the entropy of one-dimensional Gaussian distribution depends on its standard deviation S: if S is relative small, then the entropy is small, and vice versa.

On the other hand, *Information Gain* (over nominal data, with nominal classes $c_1, \ldots, c_n$) tell us how much information we obtain if consider some particular attribute:

$$I_n(x) = H_n(c_1, \ldots, c_n) - \{\ Q_1/T\ H_n(c_1) + \ldots + Qn/T\ H_n(c_n)\ \} \tag{4}$$

where Qn/T is the weight (instances quantity) for class n, respect to the total instance quantity T.

So, if we combine these ideas, we obtain a new form of Information Gain I_c applied to continuous data. For a database with continuous attributes and n nominal classes we propose the next equation (where S^2 means variance):

$$I_c(x) = S^2(c_1, \ldots, c_n) - \{\ (Q_1/T)\ S^2(c_1) + \ldots + (Q_n/T)\ S^2(c_n)\ \} \tag{5}$$

With equation (5) we can obtain feature relevance in a filter-ranking fashion, without requiring parameters' adjustments. For example, if we have a continue attribute At1:

At1 Values	0.8	0.7	0.8	0.6	0.2	0.1	0.3	0.1	0.3
At1 Class	c_1	C_1	C_1	c_1	c_2	C_2	c_2	c_2	c_2

Then $S^2(c_1, \ldots, c_n) = S^2(0.8,0.7,0.8,0.6,0.2,0.1,0.3,0.1\ 0.3) = 0.085$, and too $S^2(c_1) = S^2(0.8,0.7,0.8,0.6)=0.009$, and $S^2(c_2) = S^2(0.2,0.1,0.3,0.1,0.3)=0.01$.

Applying equation (5): $I_c(At1) = 0.085 - \{ 4/9 * 0.009 + 5/9 * 0.01 \} = 0.075$. If we repeat the process for more attributes and we obtain that: $I_c(At2) = 0.003$, $I_c(At3) = 0.28$, $I_c(At4) = 0.04$, then the attribute ranking is At3, At1, At4, At2, where At3 is the best attribute, and so on.

By analogy, we call our method for feature selection as *Variance Gain* (*vG*). Thus, the proposed method consist of:

1. Perform data normalization[1], between 0 and 1 (to maintain the same scale for all database continuous attributes).
2. Apply *vG* to each attribute (to obtain the relevance for each one).
3. Realize a descending ordering attribute-metric (attribute ranking process).
4. Select the best attributes (to select the best ranking attributes, we use a threshold defined by the largest gap between two consecutive ranked attributes, e.g., a gap greater than the average gap among all the gaps, according to [4]).
5. Use the selected attributes to perform induction (data mining process).

vG uses a simple metric based on testing decreasing values of variance in the class after selecting an attribute and results useful for the feature selection task. As shown in Section 3, *vG* is also a very effective and competitive alternative.

2.2 *dG* Method: When There Are Interdependencies Among the Attributes

The proposed method to realize non-myopic [5] feature selection is inspired also in the Shannon's (n-dimensional) entropy and the Information Gain measures. So, in analogous way, we begin our description introducing related basic concepts.

Formally, the entropy H of a numerical, or continuous distribution with an n-dimensional distribution $p(x_1, x_2, \ldots, x_n)$ has been defined [6] as:

$$H = -\int\ldots\int p(x_1, x_2, \ldots, x_n) \log_2 p(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n \quad (6)$$

Generally $p(x_1, x_2, \ldots, x_n)$ is unknown, and must be estimated, but this process has exponential complexity [7]. Again, following to Shannon, physicists and statisticians, we assume that this density distribution is Gaussian. If we realize some algebraic manipulations over (6), assuming the n-dimensional Gaussian distribution with associated quadratic form a_{ij} we obtain:

$$H = \log_2 \{ 2 \pi e \}^{n/2} |a_{ij}|^{-1/2} \quad (7)$$

where $|a_{ij}|$ is the determinant whose elements are the covariance matrix a_{ij}.

Observing (7) we can say that, for this case, the n-dimensional entropy depends on the covariance matrix determinant. Indeed, the covariance matrix is the natural generalization to higher dimensions of the concept of the variance of a scalar-valued random variable. Also we point out that the geometric meaning of a determinant is just the volume of the n-dimensional parallelepiped that n vectors (in our case, n attributes) forms. This means that, while more volume, the vectors are more independents and therefore, a larger n-dimensional entropy is obtained, and vice-versa: if $|a_{ij}|$ is relative small, then the n-dimensional entropy is small, indicating us that the features are inter-dependents.

[1] One can easily scale it so the variance is 1 and mean is 0, which is also popular.

So again, we propose combine ideas from equations (4) and (7) to obtain a new form of a non-myopic Information Gain: *dG* (determinant Gain) applied to continuous and n-dimensional data. Then, for a database with continuous attributes and discrete multi-class we propose the next equation:

$$dG = |a_{ij}|(c_1,\ldots,c_n) - \{(Q_1/T)\,|a_{ij}|(c_1) + \ldots + (Q_n/T)\,|a_{ij}|(c_n)\} \quad (8)$$

With equation (8) we can obtain a measure of subset feature relevance, without tuning of parameters and in a very simple way. For instance, we can consider the well-known XOR problem: applying the proposed metric we obtain the following evaluations for attributes *X* and *Y*: $dG(X) = 0 - (0 + 0) = 0.0$; $dG(Y) = 0 - (0 + 0) = 0.0$; $dG(X,Y) = 0.0625 - (0 + 0) = 0.0625$.

These results imply that if we consider *X* or *Y* in an isolated fashion, they cannot predict the class. On the other hand, if we evaluated the joint contribution of *X* and *Y*, the *dG* value is greater, and therefore implies that this subset is a better predictor of the class. In order to find the best (or near best) attribute subset we can apply *best-first search*. Thus, the non-myopic feature selection method proposed consist of:

Given a dataset with *N* attributes and *M* instances (where $\|\ .\ \|$ is the cardinal of a set):

1. Perform data normalization, between 0 and 1 (to maintain the same scale for all database continuous attributes);
2. Apply *dG* for each attribute;
3. While (available memory) or (unexplored nodes) do
   begin
   select for expansion the feature subset F with the best *dG* (and better than his parent node);
   for I := 1 to $(N - \|F\|)$ do
   begin
   obtain $dG\ (F \cup I \mid I \notin F)$;
   end;
   end;
4. Use the best evaluated attribute subset to perform induction (data mining process).

The proposed method *dG* is tested in the next Section.

3 Experiments

We conducted several experiments with real and synthetic datasets to empirically evaluate if *vG* and *dG* can do better in selecting features than other well-known feature selection algorithms, in terms of learning accuracy, attribute reduction and processing time. We also choose synthetic datasets in our experiments because the relevant features of these datasets are known beforehand.

3.1 Experimentation Details

The experimentation objective is to observe *vG* and *dG* behavior related to classification quality (predictive accuracy), attribute reduction and response time.

First, we test ours proposed methods with a real database with 24 attributes and 35,983 instances; this database contains information of Mexican electric billing customers, where we expect to obtain patterns of behavior of illicit customers.

We test too with five well-known databases, taken form the UCI repository [8] (see Table 2 for details).

To obtain additional evidence, we experiment with the corrAL (and corrAL-47) synthetic dataset, proposed in [4], that has four relevant attributes (A0, A1, B0, B1), plus one irrelevant (I) and one redundant (R) attributes; the class attribute is then defined by the function $Y = (A0 \wedge A1) \vee (B0 \wedge B1)$.

In order to compare the results obtained with *vG* and *dG*, we use Weka´s [9] implementation of ReliefF, CFS, OneR and ChiSquared feature selection algorithms. These implementations were run using Weka´s default values, except for ReliefF, where we define to 5 the number of neighborhood, for a more efficient response time. Additionally, we experiment with 7 Elvira´s [10] filter-ranking methods: Bhattacharyya, Matusita, Euclidean, Mutual Information, Shannon entropy, Kullback-Leibler 1 and 2. All the experiments were executed in a personal computer with a Pentium 4 processor, 1.5 GHz, and 250 Mbytes in RAM. In the following Section the obtained results are shown.

Table 1. J4.8´s accuracies for 10-fold-cross validation using the features selected by each method (Electric billing database)

Method	Total features selected	Accuracy (%)	Pre-processing time
CFS	1	90.18	9 secs.
dG	2	90.70	43 secs.
vG	3	94.02	0.7 secs.
Bhattacharyya	3	90.21	6 secs.
Matusita distance	3	90.21	5 secs.
ReliefF	4	93.89	14.3 mins.
Euclidean distance	4	93.89	5 secs.
Kullback-Leibler 1	4	90.10	6 secs.
Mutual Information	4	90.10	4 secs.
Kullback-Leibler 2	9	97.50	6 secs.
OneR	9	95.95	41 secs.
Shannon entropy	18	93.71	4 secs.
ChiSquared	20	97.18	9 secs.
All attributes	24	97.25	0

3.2 Experimental Results

Testing over the Mexican electric billing database, we use the selected features for each method as input to the decision tree induction algorithm J4.8 included in the Weka tool (J4.8 is the last version of C4.5, which is one of the best-known induction algorithms used in data mining). We notice that *dG* obtains good accuracy with only 2 attributes, better than other methods that select 3 and 4 attributes (Table 1). On the other hand, *dG* is faster than ReliefF, although this method obtains better accuracy, but selecting more attributes (4 attributes). We notice too that *vG* obtains an excellent accuracy with 3 features and it has the best processing time.

To have a better idea of the *vG* and *dG* performance, we can compare the results presented previously against the results produced by an exhaustive wrapper approach. In this case, we can calculate that, if the average time required to obtain a tree using J4.8 is 1.1 seconds, and if we multiply this by all the possible attribute combinations, then we will obtain that 12.5 days, theoretically, would be required to conclude such a process.

Testing over five UCI datasets, *vG* and *dG* obtains similar average accuracy as CFS and ReliefF, but in general with less processing time and better feature reduction than ReliefF (Table 2). SOAP's results were taken from [11]: although this method is very fast, it cannot reduce considerably the quantity of attributes for Ionosphere dataset.

Table 2. J4.8´s accuracies using the features selected by each method for five UCI datasets

Method	Autos (25/205/7)			Horse-c (27/368/2)			Hypothyroid (29/3772/4)			Sonar (60/208/2)			Ionosphere (34/351/2)			**Avg. Acc**
	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	TF	Ac	Pt	
All atts	25	*82*	0	27	*66*	0	29	*99*	0	60	*74*	0	34	*91*	0	***82.4***
vG	8	*75*	0.01	3	*69*	0.02	4	*95*	0.2	11	*73*	0.03	4	*91*	0.3	***80.6***
dG	7	*75*	12	2	*68*	14	5	*95*	26	9	*75*	14	3	*88*	18	***80.2***
CFS	6	*74*	0.05	2	*66*	0.04	2	*96*	0.3	18	*74*	0.09	8	*90*	3	***80.0***
ReliefF	11	*74*	0.4	3	*66*	0.9	6	*93*	95	4	*70*	0.9	6	*93*	4	***79.2***
SOAP	3	*73*	0.01	3	*66*	0.02	2	*95*	0.2	3	*70*	0.02	31	*90*	0.01	***78.8***
Mutual I	3	*72*	0.9	4	*68*	1	2	*90*	1.4	18	*73*	1	3	*86*	1	***77.8***
OneR	5	*70*	0.8	3	*67*	1	3	*88*	1.3	12	*72*	1	4	*85*	1	***76.4***
KL-1	3	*71*	0.9	4	*61*	1.2	3	*92*	1.7	16	*70*	1	2	*86*	1	***76.0***
KL-2	4	*68*	0.9	4	*62*	1.1	2	*89*	1.5	11	*68*	1	3	*83*	1	***74.0***
Matusita	3	*66*	1.7	3	*61*	2.3	2	*91*	3.3	17	*68*	2.5	2	*83*	2	***73.8***
Bhattac	3	*67*	0.8	3	*60*	1	1	*90*	1.4	9	*68*	1	2	*83*	1	***73.6***
Euclidea	2	*66*	1	3	*62*	1.4	2	*90*	1.2	10	*67*	1.1	2	*82*	1	***73.4***
ChiSqua	3	*67*	1	2	*60*	1.6	3	*88*	1.3	11	*65*	1.2	2	*80*	1	***72.0***
Shannon	4	*66*	0.9	4	*61*	1.3	2	*87*	1.6	9	*66*	1	2	*80*	1	***72.0***

"(25/205/7)" means (attributes, instances, classes) for Autos dataset, and so on.
TF=Total features selected Ac=Accuracy (%) Pt=Pre-processing time (secs.)

Finally, when we test with the corrAL and corrAL-47 datasets [4], our *dG* method produces the best results (Table 3) because it selects the perfect attributes, it is to say, it can detect effectively the important ones (A0, A1, B0 and B1); although *vG* cannot obtain the perfect quantity of attributes, it can detect the important ones; results for FCBF, CFS and Focus methods was taken from [4]. Elvira´s ranking methods obtain poor results, so we prefer instead show results for Symmetrical Uncertainty (SU) and Gain Ratio metrics.

Table 3. Features selected by different methods (corrAL and corrAL-47 datasets)

Method	CorrAL	corrAL-47
dG	A0, A1, B0, B1	A0, A1, B0, B1
vG	R, A0, A1, B0, B1	R,B1,$B1_1$,A1,$A1_1$,A0,$A0_1$,B0
ReliefF	R, A0, A1, B0, B1	R,$B1_1$,A0,$A0_0$,B1,$B1_0$,B0,$B0_0$,$B0_2$,A1,$A1_0$
$FCBF_{(log)}$	R, A0	R, A0, A1, B0, B1
$FCBF_{(0)}$	R, A0, A1, B0, B1	R, A0, A1, B0, B1
CFS	A0, A1, B0, B1, R	A0, A1, B0, B1, R
Focus	R	A0, A1, $A1_2$, B0, B1, R
SU (Weka)	R, A1, A0, B0, B1	$A0_1$, A0, $A0_7$, B01, B0, $A1_1$, A1, R
Gain Ratio (Weka)	R, A1, A0, B0, B1	$A0_1$,A0,$A0_7$,B0,$B0_1$, A1, R, $A1_1$
OneR	R, A1, A0, B0, B1	$A0_1$,A0,$A0_7$,$B0_1$,B0, $A1_1$, A1, R, $A0_5$, $B1_3$
ChiSquared	R, A1, A0, B0, B1	$A0_1$,A0,$A0_7$, $B0_1$,B0, $A1_1$, R, A1, $B1_3$

We point out that these results suggest that *dG* method effectively captures inter-dependencies among attributes and therefore, it is a non-myopic feature selection method.

Pre-processing time for *dG* method was inferior to one second when experiment with the corrAL dataset, and with the corrAL-47 dataset this time was 59 seconds (due to combinatorial search): we think that this response time is reasonable, but we recognize that the other ten tested methods only need around 10 seconds to conclude this task.

4 Discussion and Related Work

There is a great variety of feature selection filter methods for nominal data. Some authors consider the ID3 algorithm [12] as one of the first proposed approaches to filter (in a embedded way). Although some ID3´s extensions (like C4.5 and J4.8) manages continuous data, they perform a kind of internal binary split discretization so, these extensions, do not operate directly over continuous data.

Among the pioneering filter methods, and very much cited, are Focus [13] (that makes an exhaustive search of all the possible attribute subsets, but this is only appropriate for problems with few attributes), and Relief [14] and ReliefF (that has the disadvantage of not being able to detect redundant attributes, and also it is time consuming).

Koller [15] uses a distance metric called cross-entropy or KL-distance, that compares two probability distributions and indicates the error, or distances, among them, plus a Markov Blanket, and obtains around 50% reduction on the number of attributes, maintaining the quality of classification and being able to significantly reduce processing times (for example, from 15 hours of a wrapper scheme application, to 15 minutes for the proposed algorithm). The final result is "sub optimal" because it assumes independence between attributes, which it is not always true.

Piramuthu [3] evaluates 10 different measures for the attribute-class distance, using Sequential Forward Search (SFS), that includes the best attributes selected by each measure into a subset, such that the final result is a better attribute subset than the individual groups proposed by each method. However, the results are not compared with the original complete attribute set and so, it is not possible to conclude anything about the effectiveness of each measure; although SFS manages to reduce the search space, multiple mining algorithm runs, varying the attribute subsets, are necessary to validate the scheme and this is computationally expensive.

SOAP is a method that operates on numerical attributes and discrete or nominal class [11] and has a low computational cost: it counts the number of times the class value changes with respect to an attribute whose values have been sorted into ascending order. SOAP reduces the number of attributes as compared to other methods; nevertheless, the user has to supply the number of attributes that will be used in the final subset. This is a common problem with the *filter-ranking* methods, that output a ordered list of all attributes, according to its relevance.

In the scenario with pure continuous data, we can apply Regression Tress [16]: this method determines relevant attributes by means of co-variance between each continuous attribute and the continuous class.

Molina [17] tried to characterize 10 different feature selection methods by measuring the impact of redundant and irrelevant attributes, as well as of the number of instances. Significant differences could not be obtained, and it was observed that, in general, the results of the different methods depended on the data being used.

Perner and Apté [5] realized an empirical evaluation of feature selection based on a real-world data set, applying the CM feature subset selection method: they showed that accuracy of the C4.5 classifier could be improved with an appropriate feature pre-selection phase that at the same time reduces attribute quantity; however, due to the paper's goal, they did not realized further experiments to emphasize the CM's response time, attribute reduction or the CM's ability to detect attribute interactions.

Other proposals for feature selection explore the use of neural networks, fuzzy logic, genetic algorithms, and support vector machines [1], but they are computationally expensive and have one, or more, user's parameters to adjust. In general, it is observed that the methods that have been proposed: a) need nominal data; b) obtain results that vary with the domain of the application; c) obtain greater quality results, only with greater computational cost; d) depend on suitable tuning; and e) they suffer of myopic feature selection.

5 Conclusions and Future Work

We have presented two feature selection new methods easy to implement that try to overcome some drawbacks found with traditional pure nominal or discrete feature selection methods.

From the experimentations presented, with a real Mexican electric billing database, five UCI datasets and two synthetic datasets, the proposed methods *vG* and *dG* represents a promising alternatives, compared to other methods, because of its acceptable processing time and good performance in the feature selection task in both, accuracy and attribute reduction. Additionally, ours methods works without user parameters that generally imply some kind of special and time consuming tuning.

Some future research issues arise with respect to *vG* and *dG* testing and improvement. For example: experimenting with more real and challenging databases (e.g., future work will be the application of the formalism to other very large power system databases such as the national power generation performance database, the national transmission energy control databases, the de-regulated energy market database, and the Mexican electric energy distribution database); applying other data mining induction-classification algorithms (e.g., Naïve Bayes classifier, 1NN, etc.); perform more experiments following [18]; apply statistical tests to observe if the differences in accuracies of the proposed methods are significant and apply other functions to overcome the Gaussian distribution assumption.

References

1. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research 3, 1157–1182 (2003)
2. Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial Intelligence Journal, Special issue on relevance, 273–324 (1997)
3. Piramuthu, S.: Evaluating feature selection methods for learning in data mining applications. In: Proc. 31st annual Hawaii Int. conf. on system sciences, pp. 294–301 (1998)
4. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5, 1205–1224 (2004)
5. Perner, P., Apté, C.: Empirical Evaluation of Feature Subset Selection Based on a Real-World Data Set. In: Zighed, A.D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 575–580. Springer, Heidelberg (2000)
6. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal 27, 379–423, 623–656 (1948)
7. Miller, E.: A new class of entropy estimators for multi-dimensional densities. In: Int.conference on Acoustics, Speech and Signal Processing (2003)
8. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases, Department of Information and Computer Science. University of California, Irvine, CA (1998), http://www.ics.uci.edu/ mlearn/MLRepository.html
9. www.cs.waikato.ac.nz/ml/weka (2004)
10. www.ia.uned.es/ elvira/ (2004)
11. Ruiz, R., Aguilar, J., Riquelme, J., SOAP: efficient feature selection of numeric attributes, VIII Iberamia, workshop de minería de datos y aprendizaje, Spain, 2002, pp. 233-242.
12. Quinlan, J.: Unknown attribute values in ID3. In: Int. conf. Machine learning, pp. 164-168 (1989)
13. Almuallim, H., Dietterich, T.: Learning with many irrelevant features. In: Ninth nat. conf. on AI, pp. 547–552. MIT Press, Cambridge (1991)
14. Kira, K., Rendell, L.: The feature selection problem: traditional methods and a new algorithm. In: Tenth nat. conf. on AI, pp. 129–134. MIT Press, Cambridge (1992)

15. Koller, D., Sahami, M.: Toward optimal feature selection. In: Int. conf. on machine learning, pp. 284–292 (1996)
16. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees. Wadsworth International Group, Belmont, CA (1984)
17. Molina, L., Belanche, L., Nebot, A.: Feature selection algorithms, a survey and experimental eval. In: IEEE Int.conf.data mining, Maebashi City Japan, pp. 306–313. IEEE Computer Society Press, Los Alamitos (2002)
18. Ambroise, C., McLachlan, G.J.: Selection Bias in Gene Extraction in Tumour Classification on Basis of Microarray Gene Expression Data. PNAS 99(10), 6562–6566 (2002)