World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:1, No:8, 2007

Attribute Selection Methods Comparison for Classification of Diffuse Large B-Cell Lymphoma

Helyane Bronoski Borges, and Júlio Cesar Nievola

***Abstract*—**The most important subtype of non-Hodgkin's lymphoma is the Diffuse Large B-Cell Lymphoma. Approximately 40% of the patients suffering from it respond well to therapy, whereas the remainder needs a more aggressive treatment, in order to better their chances of survival. Data Mining techniques have helped to identify the class of the lymphoma in an efficient manner. Despite that, thousands of genes should be processed to obtain the results. This paper presents a comparison of the use of various attribute selection methods aiming to reduce the number of genes to be searched, looking for a more effective procedure as a whole.

***Keywords*—**Attribute selection, data mining.

I. Introduction

ANALYSIS of gene expression is of the foremost importance to Biology. This kind of analysis can give important insights about a cell's function, since changes in the physiology of an organism are generally followed by changes in genes expression patterns [1]. Using DNA micro-array technique, it became possible to create a systematic categorization of gene expression in malignant B cells in Diffuse Large B-Cell Lymphoma (LDGCB).

LDGCB is the most common subtype of non-Hodgkin lymphoma and, despite many patients with this disease can become healthy again by chemotherapeutic combinations; a large group of patients remain not healthy. Through the study of these groups, two distinct kinds of LDGCB cells were identified. They had their gene expression patterns presented by two different stages in B cell's differentiation. The first kind of gene expression characterizes the germinative center of B cells and the second one is normally induced when B cells are activated [2].

Recently, Data Mining techniques started to be used in the analysis of data obtained from micro-arrays, both for classification and for clustering of the resulting data. It was noticed that the analysis of this kind of data has some particularities regarding general data, since the number of instances (sample size) is small (in the tens or few hundreds), relative to the large number of attributes (which correspond to each gene that has its behavior represented in each sample – typically in the range of thousands of genes).

Clustering techniques (especially hierarchical ones) are widespread in Biology, where clusters of data with similar behavior are searched. It's done in the hope that these clusters represent situations where the elements have interesting emergent properties. Classification techniques, on their own, are considered more useful in a second round, after the discovery of the subjacent clusters, especially in clinical analysis with the purpose of diagnostic or prognostic.

Authors are with PPGIA - Pontifícia Universidade Católica do Paraná (PUCPR), Caixa Postal 16.210 – 81.611-970 – Curitiba – PR – Brasil (e-mail: {helyane, nievola} @ppgia.pucpr.br).

II. Problem Description

Lymphomas are lymphatic system's cancer (malignant tumor). Since they were first observed by Thomas Hodgkin in 1832 [7], these diseases has been widely studied, and many classification systems have been proposed. Lymphomas' complexity and heterogeneity are related to the diversity of lymphoid tissue's cells. Therefore, in the last classification of the lymphoid and hematopoietic tissues' tumors, the use of clinical aspects, morphology, imunophenotypage and molecular analysis were commended, as often as possible, to diagnose these malignancies.

Human lymphoma classification has evolved since its initial recognition, starting with the distinction of Hodgkin's disease from other malignant and non-malignant conditions. The term Hodgkin's disease was proposed in 1865 by Wilks, based in Thomas Hodgkin's initial observations. Later on, Bilbroth has proposed the term malignant lymphoma for these lesions' category, since traditionally the "oma" suffix is used to represent a group of malignancies of the lymphoid tissue with specific microscopic characteristics, while other malignancies that don't present these characteristics were named non-Hodgkin's lymphomas [4].

Many different classifications of the lymphomas have been proposed based on morphological and molecular parameters. Two of the most recent classification's schemas of lymphoid tissue malignancies take into account the imunophenotype of each entity. Despite that, in these classification's schemas many morphological subtypes are unified in clusters, though they are believed to "include more than one disease entity".

It is important to differentiate between Hodgkin's disease and non-Hodgkin lymphoma, because the treatment of the last one should be more aggressive. Clinically, patients with Hodgkin's disease present disease progression by adjacency and they have, generally, a more homogeneous and dragged evolution; patients with non-Hodgkin's disease on the other hand can have more indolent, aggressive or highly aggressive behavior, depending on the type. Diffuse large B cell lymphomas (LGCBD) constitute a group of lymphomas that include different subtypes, with clinical and histological variables. They are characterized as being aggressive and malignant lymphomas, with annual incidence of 25.000 cases, and they have approximately 40% of the non-Hodgkin's lymphoma. Even though many patients with diffuse large B cell lymphoma (LGCBD) could respond well to therapy by

chemotherapeutic combinations, a large number of them remain ill. In recent years many risk factors were introduced in these lymphomas' evaluation, in an attempt to better determine the groups of bad prognosis and to delineate more efficient therapeutic procedures.

Using DNA micro-array technique and non-supervised learning, with hierarchical clustering, Alizadeh and her colleagues have identified two molecularly distinct forms of diffuse large B cell lymphomas (LGCBD), which had gene expression patterns indicative of different stages of B cell differentiation. One type expresses genes characteristics of germinal center B cells and the other type expresses genes normally induced during in vitro activation of peripheral blood B cells [2].

In 2002, given sequence to the study of diffuse large B cell lymphoma (LGCBD) initiated by Alizadeh and her colleagues, Shipp used supervised learning to classify the disease in LDGCB and Follicular Lymphoma (FL), and to identify cured versus fatal or refractory disease after chemotherapy, in databases other than those used in the previous study [13].

III. Data Mining Process

When genetic expression profiles are studied, unknown data are manipulated very often. The data can be redundant and even, sometimes, irrelevant. In order to reduce the problems associated with such characteristics, there exists an initial stage in data mining called preprocessing, which tries to attain a better data quality and as a result improving the results of the mining algorithm [10]. To take it into account, the following data mining phases were executed: data consolidation, preprocessing, data mining and post-processing.

A. Data Consolidation

The data used in the experimentation were the same as used by Alizadeh and her colleagues. They were downloaded from a public domain repository of biomedical data. This database uses data from genetic expression data, originated from micro-array DNA technique, which allows the measure the expression level of thousands of genes in one single experiment.

The dataset employed consists of 47 samples, 24 of them belonging to the germinative center of B cell group, while 23 belong to the activation B cell group. Each sample is represented by 4026 genes, all of them through its numeric value. The 4027^{th} value is the goal attribute. The goal is to identify to which class each sample is related: *germinal* or *activated*.

B. Preprocessing

A very important step in the data mining process is data preprocessing. This stage is responsible for consolidation of relevant information to the mining algorithm, trying to reduce the problem complexity. Among the steps in preprocessing, attribute selection has a special role.

Attribute selection is a process in which a subset of M attributes out of N is chosen, complying with the constraint $M \leq N$, in such a way that characteristic space is reduced according to some criterion [10]. Attribute selection guarantees that data getting to the mining phase are of good quality [10].

Algorithms used for attribute selection can be normally separated in two main activities: search for the attributes subset and evaluation of the subsets found, as can be seen in Fig. 1.

Search algorithms used in the first stage can be subdivided in 3 main groups: exponential, random and sequential algorithms [5]. Exponential algorithms, as for instance the exhaustive search, try all possible attribute combinations before returning the attribute subset. Normally, they are not computationally feasible, since their running time grows exponentially in the number of available attributes [10]. Genetic algorithms are one example of random search methods, and their main advantage over sequential ones is that they are capable of dealing with the problem of attribute interaction [6].

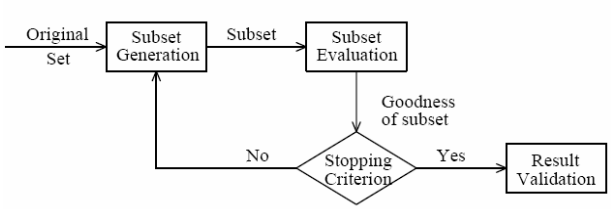

Fig. 1 Steps in Attribute Selection [11]

Sequential algorithms are relatively efficient in the solution of many attribute selection problems; despite they have the disadvantage of not taking attribute interaction into account. Two examples of sequential algorithms are forward selection and backward elimination.

Sequential forward selection starts the search for the best attribute subset with an empty set of attributes. Initially, attribute subsets with only one attribute are evaluated, and the best attribute A^* is selected. This attribute A^* is then combined with all other available attributes (pairwise), and the best subset of two attributes is selected. The search goes on with this procedure, incorporating one attribute at a time to the best attribute subset already selected, until the quality of the best selected attribute subset cannot be further improved.

Contrary to forward selection, sequential backward elimination starts the search for the best attribute subset with a solution representing all attributes, and at each iteration one attribute is removed from the actual solution, until no further improvement in the quality of the solution can be attained.

Regarding the evaluation of the generated attribute subsets, two main approaches can be implemented: filter approach or wrapper approach. Both approaches are independent from the algorithm used for the selection of the candidate subsets, and they are characterized by their degree of dependence regarding the classification algorithm.

The wrapper approach defines an adequate subset of solutions to a previous chosen database and a particular induction algorithm, taking into account the inductive bias of the algorithm and its interaction with the training set. Fig. 2 represents an attribute selection algorithm that uses the wrapper approach.

Different from the wrapper approach, the filter approach tries to chosen an attribute subset independently from the classification algorithm to be used, making an estimate of

attribute quality looking just to the data. Fig. 3 presents the schema of attribute selection with a filter approach that makes the selection using a preprocessing step based only on the training data. During this phase, the generated attribute sets can be evaluated according to some simple heuristic, as, for instance, the ortogonality of the data [3].

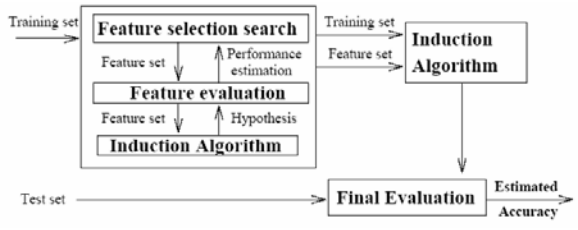

Fig. 2 Attribute selection using the wrapper approach [9]

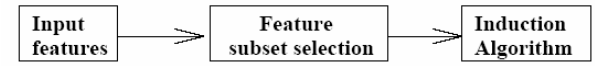

Fig. 3 Attribute selection with filter approach

Normally, the wrapper approach has a large algorithm running time, but the number of correctly classified instances tends to be greater than that obtained by the filter approach. There are many techniques to evaluate an attribute subset with the filter approach. Among the evaluation measures some deserve attention, as Relevance and Consistency [12]. Relevance measure quantifies how much two attributes are associated, that is to say, whether it is possible to predict some attribute's values, when some other attribute's value is know. Within the attribute selection context, the best evaluated attribute is the one that best predicts the class.

By using consistency, the evaluation of attributes subset tries to determine the class' consistency level when the training instances are projected onto the attributes subset. For evaluation purposes, using the wrapper approach, the following learning algorithms were used: C4.5, Bayesian net, Naive Bayes, k-NN (with *k* taking the values *k*=1, *k*=3, *k*=5 e *k*=7) and Decision Table. The corresponding classifier algorithm used in the selection was used to classify each attributes subset.

C. Data Mining

This is the most important step in the knowledge discovery process. It is defined as the search for interesting relationships and patterns that exist in real world databases, but that are hidden among a huge amount of stored data. These relationships represent valuable knowledge about the database and, consequently, about the real word domain they represent [8]. Among the main tasks possible for the Knowledge Discovery, classification was chosen. The main goal of classification is to discover the relationship between the predictor attributes and the goal attribute. The data used in the experimentation were extracted from the *Kent Ridge*[1] biomedical data repository, publicly available. This base is identified as LDGCB and is available in C4.5 format, which is composed by two files named *.data* e *.names*.

Using the above mentioned approaches, various experiments were made. They are listed in Table I.

TABLE I
LIST OF EXPERIMENTS AND THEIR RESULT

Attribute Selection			
Experiment	**Search Method**	**Subset Evaluation**	**Number of attributes selected**
1	Genetic search	Consistency	305
2	Forward selection	Consistency	3
3	Genetic search	Relevance	1769
4	Forward selection	Relevance	33
5	Forward selection	Wrapper with Bayesian Net	2
6	Forward selection	Wrapper with Naïve Bayes	3
7	Forward selection	Wrapper with C4.5	1
8	Forward selection	Wrapper with k-NN (k=1)	3
9	Forward selection	Wrapper with k-NN (k=3)	2
10	Forward selection	Wrapper with k-NN (k=5)	3
11	Forward selection	Wrapper with k-NN (k=7)	4
12	Forward selection	Wrapper with D.T.[2]	3
13	Genetic search	Wrapper with C4.5	892
14	Genetic search	Wrapper with Bayesian net	1367
15	Genetic search	Wrapper with Naïve Bayes	1391
16	Genetic search	Wrapper with D.T.[2]	1173
17	Genetic search	Wrapper with k-NN (k=1)	2035
18	Genetic search	Wrapper with k-NN (k=3)	1411
19	Genetic search	Wrapper with k-NN (k=5)	1058
20	Genetic search	Wrapper with k-NN (k=7)	1940

IV. RESULTS AND DISCUSSION

Table II shows the results obtained when all attributes are used to build each one of the classifiers used in the experiments: Decision tree (built up by C4.5 algorithm), Bayesian net, *Naive Bayes*, k-NN for k=1, k=3, k=5 e k=7 and Decision Table. All the results presented there were achieved

[1] http://sdmc.lit.org.sg/GEDatasets/Datasets.html

[2] Decision Table.

by using cross-validation (factor 10) with a p-value of 0.05. The value should be read as median ± standard deviation.

Tables III, IV and V show the results achieved by each classifier when using the restricted subset of attributes that means, through the use of only those attributes that were selected by each configuration of search method and evaluation method of the generated subsets. These results present a large span, ranging from situations were 2035 attributes (genes) were selected (in the case of genetic search as the search method with evaluation of generated subsets using the nearest neighbor algorithm - only one neighbor, k=1), until the extreme situation where only one was selected during the attribute selection process (in the case of the combination of forward selection as the search method with the wrapper approach for the evaluation of the generated subsets, employing the decision tree algorithm - C4.5). All the results presented there were achieved by using cross-validation (factor 10) with a p-value of 0.05. The value should be read as median ± standard deviation.

When looking at the results, one characteristic that presents itself immediately is the greater span in the number of selected attributes, according to the search method and the evaluation approach of each subset. When looking deeper in the results at the selected attributes (genes) (not showed in the tables) the gene known as GENE3330X stands out: it was selected by almost all of the selection procedures. Therefore, it seems clear the importance of this specific gene in the classification of the kind of lymphoma.

Making a comparison of the achieved results, it is pretty evident that attribute selection hás got better classification results in almost every situation. The use of the wrapper approach for evaluation purposes has lead to better results in a consistent way. This result, as expected, was evident, since the classifier was chosen consistently with the algorithm used during the attribute selection stage.

V. Conclusion

Attribute selection has decreased significantly data dimensionality, leading to a better performance of the data mining algorithm, resulting in inferior running time compared to the situation where all the attributes of the original database were used. Besides that, experiments verified that subsets generated through the selection attribute the number of correctly classified instances is higher, sometimes achieving 100%.

TABLE II
Proportion of Correctly Classified Instances with all Attributes being used

C4.5	Bayesian Net	Naive Bayes	k-NN				Decision Table
			k = 1	k = 3	k = 5	k = 7	
77% ± 23.7%	97.5% ± 7.9%	97.5% ± 7.9%	75.5% ± 21.2%	77% ± 17.5%	75% ± 23.6%	73% ± 18.7	84% ± 23.3%

TABLE III
Proportion of Correctly Classified Instances in the Selected Attributes Subset by Genetic Search and Forward Selection using Relevance and Consistence for the Evaluation (In %)

Experiment	C4.5	Bayesian Net	Naive Bayes	k-NN				Decision Table
				k = 1	k = 3	k = 5	k = 7	
1	83 ± 13.1	78.5 ± 13.5	89 ± 15.6	72.5 ± 8.9	66 ± 16.9	75 ± 15.6	72.5 ± 19.9	55.5 ± 17
2	89 ± 11.7	93.5 ± 10.5	94 ± 9.6	93.5 ± 10.5	96 ± 8.4	93.5 ± 10.5	93.5 ± 10.5	98 ± 6.3
3	80 ± 16.8	100 ± 0	97.5 ± 7.9	78 ± 15.3	85.5 ± 13.8	78.5 ± 22.3	73 ± 18.7	86.5 ± 16.6
4	76.5 ± 30	100 ± 0	100	98 ± 6.3	98 ± 6.3	96 ± 8.4	96 ± 8.4	88.5 ± 17

TABLE IV
Proportion of Correctly Classified Instances in the Selected Attributes Subset by Forward Selection Search using Wrapper Approach

Forward Selection search plus Wrapper Approach		
Experiment	Classifier	Correctly classified instances (%)
5	Bayesian net	98.0 ± 6.3
6	*Naive Bayes*	98.0 ± 6.3
7	C4.5	91.5 ±11.0
8	k-NN with k=1	98.0 ± 6.3
9	k-NN with k=3	98.0 ± 6.3
10	k-NN with k=5	98.0 ± 6.3
11	k-NN with k=7	100.0 ± 0.0
12	Decision Table	98.0 ± 6.3

TABLE V
PROPORTION OF CORRECTLY CLASSIFIED INSTANCES IN THE SELECTED ATTRIBUTES SUBSET BY GENETIC ALGORITHM SEARCH USING WRAPPER APPROACH

Genetic Algorithm Search plus Wrapper Approach		
Experiment	Classifier	Correctly Classified instances (%)
13	C4.5	91.5 ± 11.0
14	Bayesian net	100.0 ± 0.0
15	*Naive Bayes*	100.0 ± 0.0
16	Decision Table	98 ± 6.3
17	k-NN with k=1	84.5 ± 14.4
18	k-NN with k=3	79.5 ± 16.4
19	k-NN with k=5	79.0 ± 20.2
20	k-NN with k=7	81.5 ± 17.6

REFERENCES

[1] Alberts, B. et al. Biologia Molecular da Célula. Editora Artes Médicas, 3ª Edição, 1997.
[2] Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 4051, 503–511 (2000).
[3] Bala, J.; Jong K. De; Huang, J.; Vafaie, H.; Wechsler, H; Using Learnig to Facilite the Evolution of Features for Recognizing Visual Concepts, In: Special Issue of Evolutionary Computatuion – Evolution, learning and Instinct: 100 years of Baldwin Effect, Vol. 4 , pp. 297-311. 1996.
[4] Billroth, T., Multiple Lymphoma: Erfolgreiche Behandlung mit Arsenik, *Deutsch Med. Wschr*, Stuttgart, V. 21, 1066-1067, 1871.
[5] Boz, O., Feature Subset Selection by Using Sorted Feature Relevance, In: ICMLA 2002 – International Conference on Machine Learning and Applications, USA, 2002.
[6] *Freitas*, A. A.; Understanding the Crucial Role of Attributes Interaction in Data Mining, In: Artificial Intelligence Review 16, pp 177-199, Kluwer Academic Publishers, 2001.
[7] Hodgkin, T., On Some Morbid Appearances of the Absorbant Glands and Spleen, *Med.-Chir. Trans.*, 17, 68-114, 1832.
[8] Holsheimer, M.; Siebes, A., Data Minig – The Search for Knowledge in Databases, Report CS-R9406, Amsterdam, 1991.
[9] Kohavi, R.; John, G. H., The Wrapper Approach, In: H. Liu & H. Motoda (Eds.) Feature Extraction, Construction and Selection: a data mining perspective, 33-49. Kluwer, 1998.
[10] Liu, H., Motoda, H., Feature Selection for Knowledge Discovery and Data Minig, Kluwer academic Publishers, 1998.
[11] Liu, H., Motoda, H., Yu, L., The Handbook of Data Mining, Lawrence Erlbaum Associates, Inc. Publishers. Editor: N. Ye. PP 409 - 423. 2003.
[12] Molina L. C., Belanche L., Nebot A. Feature Selection Algorithms: A Survey and experimental Evaluation. Technical Report LSI-02-62-R Universitat Politècnica de Catalunya, Barcelona, Spain, 2002.
[13] Shipp, M.A. et al. Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nature*, Vol. 8, N. 1, 68-74, 2002.

Helyane Bronoski Borges is a student of the Graduate Program in Computer Science at the Pontifícia Universidade Católica do Paraná, Brazil. Her research' interests include data mining, attribute selection, machine learning, and bioinformatics.

Júlio Cesar Nievola is with the Graduate Program in Computer Science at the Pontifícia Universidade Católica do Paraná, Brazil. He's the leader of the Machine Learning and Data Mining Research Group within the Institution and his research' interests include Data Mining, Attribute Selection, Machine Learning, and their applications in the Bioinformatics field.