Data Mining Approaches to Diffuse Large B–Cell Lymphoma Gene Expression Data Interpretation

Jesús S. Aguilar-Ruiz[1], Francisco Azuaje[2], and José C. Riquelme[3]

[1] University of Seville, Seville, Spain
aguilar@lsi.us.es
[2] University of Ulster, North Ireland
fj.azuaje@ulster.ac.uk
[3] University of Seville, Seville, Spain
riquelme@lsi.us.es

Abstract. This paper presents a comprehensive study of gene expression patterns originating from a diffuse large B–cell lymphoma (DLBCL) database. It focuses on the implementation of feature selection and classification techniques. Thus, it firstly tackles the identification of relevant genes for the prediction of DLBCL types. It also allows the determination of key biomarkers to differentiate two subtypes of DLBCL samples: *Activated B–Like* and *Germinal Centre B–Like DLBCL*. Decision trees provide knowledge–based models to predict types and subtypes of DLBCL. This research suggests that the data may be insufficient to accurately predict DLBCL types or even detect functionally relevant genes. However, these methods represent reliable and understandable tools to start thinking about possible interesting non–linear interdependencies.

1   Introduction

Lymphomas are divided into two general categories: Hodgkin's disease (HD) and non–Hodgkin's lymphoma (NHL). Over the past 20 years HD rates have declined, accounting for only 1% of all cancer. By contrast, NHL cases have increased by more than 50% during the same period in the United States [10]. They represent 4% of all cancer cases, becoming the fifth most common malignancy in that country. An analysis of NHL incidence trends between 1985 and 1992 in seven European countries showed an average increase of 4.2% per year, in the absence of an increase in the incidence of HD. In Spain, their death rate per 100.000 people during the periods 1965–69 and 1995–98 increased 212.7% for men and 283.9% for women [13]. These figures reveal the significance of developing advanced diagnostic and prognostic systems for these diseases.

In the last two decades, a better understanding of the immune system and the genetic abnormalities associated with NHL have led to the identification of several previously unrecognized types of lymphoma. However, this is a complex and expensive task. For instance, distinctions between Burkitt's lymphoma

and diffuse large B–cell lymphoma (DLBCL) often prove to be difficult, since High–grade B–cell Burkitt–like lymphoma appears to be clinically similar to DLBCL [17]. DLBCL is the most common type of NHL. There are no reliable morphological or immunohistochemical indicators that can be used to recognize subtypes of DLBCL. Moreover, it is fundamental to develop treatments specially tailored to the individual requirements and profiles of patients [9].

Others authors, such as Shipp et al. [16], have studied the distinction between DLBCL and a related germinal centre B–cell lymphoma, known as follicular lymphoma (FL), which over the time acquires morphological and clinical features observed in DLBCLs.

Technological advances now allow to screen the expression of thousands of genes in parallel. Thus, gene expression profiling has become crucial for the development of powerful diagnostic and prognostic methods for these types of cancers. It offers an opportunity to characterise the biological behaviour of specific tumours. Such an approach, which is based on the use of micro–array techniques, may provide massive amounts of information. Therefore, there is a need for sophisticated algorithms capable of extracting novel and useful knowledge from a biomedical point of view. In fact, gene–expression profiling is thought to be a revolutionary technique to support the diagnosis of cancers [1].

A basic approach to the study of expression data consists of applying traditional statistical techniques. In many problems these methods have shown to be unable to extract relevant knowledge from data. Thus, the Knowledge Discovery in Databases approach (KDD) [8] represents a useful framework for addressing the challenges of gene expression analysis. In particular, feature selection techniques may significantly support diagnostic studies, based on the identification of relevant genes or biomarkers. *Clustering* is another important task within KDD, which aims to organize the information in terms of their similarity patterns [6]. *Supervised classification* has also become an important goal in this type of studies [5]. Techniques such as *decision trees* [4] or *decision lists* [15] may represent more effective and understandable tools for aiding in the prediction of types or subtypes of diseases.

In this paper, we carry out a broad study of a well–known database generated by Alizadeh et al. [2], who investigated the identification of lymphomas and DLBCL subtypes based on expression patterns. The data comprise 96 samples described by the expression values of 4026 genes.

Firstly, feature selection techniques are implemented to identify relevant genes for the prediction of lymphoma cancer types. Moreover, it allows the detection of biomarkers to distinguish two subtypes of DLBCL: *Activated B–Like* and *Germinal Centre B–Like Lymphomas*. The following methods have been implemented: Information gain criterion, based on the entropy measure, the Relief method and $\chi 2$ ranking and filtering. Decision trees are constructed to perform classification tasks initially based on the original classes, and afterwards using the DLBCL subtypes.

This study reveals that not only the genes identified by Alizadeh et al. are relevant for the prediction of the two subtypes of DLBCL, but many others groups

Table 1. Most relevant genes provided by Relief, InfoGain and χ^2, ordered by ranking.

ReliefF	Freq.	InfoGain	Freq.	χ^2	Freq.
GENE1610X	•••	GENE707X	••	GENE2400X	•••
GENE1636X	•	GENE655X	•••	GENE788X	••
GENE1648X	•	GENE694X	•••	GENE3639X	••
GENE1622X	•••	GENE1622X	•••	GENE707X	••
GENE1702X	•	GENE844X	••	GENE655X	•••
GENE653X	•••	GENE1635X	••	GENE1992X	•
GENE1637X	•	GENE2400X	•••	GENE1675X	••
GENE712X	••	GENE1610X	•••	GENE694X	•••
GENE1607X	•	GENE717X	••	GENE3767X	•
GENE611X	•	GENE711X	•	GENE769X	••
GENE1647X	•	GENE639X	••	GENE2387X	••
GENE708X	••	GENE2402X	••	GENE1622X	•••
GENE1651X	•	GENE769X	••	GENE1610X	•••
GENE2402X	••	GENE641X	•	GENE2032X	•
GENE537X	•	GENE628X	•	GENE467X	••
GENE1658X	•	GENE669X	•	GENE3685X	•
GENE654X	••	GENE2403X	•••	GENE2403X	•••
GENE1608X	•	GENE647X	•	GENE1371X	•
GENE2393X	•	GENE712X	••	GENE2033X	•
GENE1641X	•	GENE783X	••	GENE646X	••
GENE721X	•	GENE653X	•••	GENE753X	•••
GENE651X	••	GENE691X	•	GENE783X	••
GENE1644X	•	GENE753X	•••	GENE764X	•
GENE1635X	••	GENE2495X	•	GENE639X	••
GENE753X	•••	GENE651X	••	GENE2428X	•

may also be considered as relevant markers. In general, KDD techniques demonstrate to be efficient for extracting valid and useful knowledge from biomedical data.

2   Feature Selection for Gene Expression Data

Feature subset selection is the process of identifying and removing irrelevant or redundant attributes. Decreasing the dimensionality of the data reduces the size of the hypothesis space and allows learning algorithms to operate faster and more effectively. It leads to smaller and easy–to–understand knowledge models of the target concept. Feature selection techniques produce ranked lists of attributes, providing the data analyst with insight into their data by clearly demonstrating the relative merit of individual attributes.

In this study, we used three feature selection techniques, both belonging to the *filter* category [7], *Information Gain Attribute Ranking*, ReliefF [11, 12] and χ^2 [14]. The information gain attribute ranking is often used where the sheer dimensionality of the data precludes more sophisticated attribute selection techniques, as the case being investigated here, which consist of 4026 attributes. ReliefF works by randomly sampling an instance from the data and then locating its nearest neighbour from the same and a different class. When dealing with noisy data the k nearest neighbours should be obtained. If the data contains multiple classes, the contributions of the k nearest neighbours can be weighted using the prior probabilities associated with each class. The values of the attributes of the nearest neighbours are compared to the sampled instance and used to up-

Table 2. Most relevant genes provided by Relief, InfoGain and χ^2, ordered by ranking.

ReliefF	Freq.	InfoGain	Freq.	χ^2	Freq.
GENE717X	• •	GENE467X	• •	GENE2202X	•
GENE2403X	• • •	GENE646X	• •	GENE2199X	• •
GENE2270X	• •	GENE3639X	• •	GENE844X	• •
GENE784X	•	GENE2395X	• •	GENE777X	• •
GENE2486X	•	GENE2668X	• •	GENE654X	• •
GENE1603X	•	GENE788X	• •	GENE1990X	•
GENE2489X	•	GENE1672X	•	GENE2424X	• •
GENE703X	•	GENE2379X	•	GENE276X	•
GENE692X	•	GENE770X	• •	GENE2862X	•
GENE2271X	•	GENE648X	•	GENE794X	•
GENE2401X	•	GENE642X	•	GENE770X	• •
GENE1653X	•	GENE593X	•	GENE768X	•
GENE1646X	•	GENE1606X	•	GENE2778X	•
GENE2244X	•	GENE734X	•	GENE3764X	•
GENE694X	• • •	GENE604X	•	GENE2395X	• •
GENE655X	• • •	GENE777X	• •	GENE2374X	• •
GENE538X	•	GENE1673X	•	GENE1324X	•
GENE731X	•	GENE2374X	• •	GENE1343X	•
GENE2668X	• •	GENE2199X	• •	GENE2795X	•
GENE584X	•	GENE649X	•	GENE653X	• • •
GENE1776X	•	GENE708X	• •	GENE1320X	•
GENE713X	•	GENE1675X	• •	GENE3334X	•
GENE2400X	• • •	GENE2387X	• •	GENE2000X	•
GENE710X	•	GENE2424X	• •	GENE473X	•
GENE714X	•	GENE706X	•	GENE1323X	•

date relevance scores for each attribute. The rationale is that a useful attribute should differentiate instances from different classes, and has the same value for instances belonging to the same class. χ^2 statistic conducts a significance test on the relationship between the values of an attribute and the classes.

This study presents results based on the original set of genes (4026), and on a subset of 50 relevant genes extracted from them, in order to compare our result to that identified by Alizadeh et al. [2].

2.1   Lymphoid Malignancies

The three methods provided different results and they were compared to find coincidences. Tables 1 and 2 show the genes selected by each method, and ordered by their relevance from top to bottom. The genes are represented with the identifiers used by Alizadeh et al. Figure 1 provides the gene names associated with each identifier. We found 8 common genes for the three methods, 25 common pairs of genes and 76 genes that were selected by only one method. Note that there are 105 different genes in Tables 1 and 2, from 150 possible selections.

Figure 1 shows these genes in terms of degrees of relevance: Very high and high relevance. *Very highly relevant* genes are those which have been selected by all of the feature selection methods. *Highly relevant* genes are those which have been identified by any two of the methods. In total, there are 8 very highly relevant genes and 25 highly relevant genes.

The next task is to know whether these 8 very highly relevant genes or, in the worst case, the 33 genes including the highly relevant ones, can predict a cancer class. This will be shown using decision trees in section 3. Alizadeh et al.

ID	Name
GENE653X	Lactate dehydrogenase A
GENE655X	GRSF-1=cytoplasmic G-rich mRNA sequence binding factor
GENE694X	Cyclin A
GENE753X	Similar to MCM2 = DNA replication licensing factor
GENE2400X	Unknown
GENE2403X	Unknown
GENE1610X	Mig=Humig=chemokine targeting T cells
GENE1622X	CD63 antigen (melanoma 1 antigen)
GENE467X	C-1-Tetrahydrofolate Synthase, cytoplasmic
GENE639X	hepatoma-derived growth factor
GENE646X	nm23-H2=NDP kinase B=Nucleoside dephophate kinase B
GENE651X	tubulin-beta
GENE654X	dystrobrevin B DTN-B2=dystrophin-associated protein A0
GENE707X	Topoisomerase II alpha (170kD)
GENE708X	Ki67 (long type)
GENE712X	Cyclin B1
GENE717X	aurora/IPL1-related kinase
GENE769X	14-3-3 epsilon
GENE770X	14-3-3 epsilon
GENE777X	semaphorin V=homologue of nerve growth cone guidance signaling proteins
GENE783X	Glycyl tRNA synthetase
GENE788X	SRPK1=serine kinase
GENE844X	ets-2=ets family transcription factor
GENE1635X	osteonectin=SPARC=basement membrane protein
GENE1675X	FCERI=Fc epsilon receptor gamma chain=High affinity immunoglobulin epsilon receptor gamma
GENE2199X	Unknown UG Hs.71252 ESTs
GENE2374X	PKC beta =Protein kinase C, beta
GENE2387X	Unknown UG Hs.181297 ESTs
GENE2395X	Unknown UG Hs.59368 ESTs
GENE2402X	Unknown
GENE2424X	Similar to neuropathy target esterase
GENE3639X	KIAA0053
GENE2668X	Mad2=MXI-1=MAX-binding protein=antagonizer of myc transcriptional activity= =Mxi-1/Max heterodimers repress c-myc targets

VERY HIGH RELEVANCE (rows GENE653X–GENE1622X); *HIGH RELEVANCE* (rows GENE467X–GENE2668X)

Fig. 1. Relevant genes to differentiate the lymphoma classes from the complete dataset (96 patients and 4026 attributes). The 50 most relevant genes for each feature selection method were selected. *very highly relevant* means that the gene was relevant for the three methods; *highly relevant* means that it was relevant for any two methods. There is no order of relevance in the list.

discovered 50 relevant genes to differentiate the GC B–Like from the Activated B–Like subtypes of DLBCL. However, only one of them, GENE2395X, has been identified in our analysis.

It is important to note that several attributes are strongly correlated with others (Pearson's correlation coefficient greater than 0.90), even among those genes selected as very highly or highly relevant in Figure 1. This fact shows that most methods for feature selection do not take into account the correlation among extracted features. Therefore, this information needs to be post–processed, removing genes of similar functionality. For instance, GENE769X and GENE770X, shown in Figure 1, or GENE1719X and GENE1720X, in Figure 2.

2.2 DLBCL Subtypes

In this section the 45 DLBCL samples are analysed. This category is divided into to subtypes: Activated B–like DLBCL (ACL) and Germinal Centre B–like

ID	Name
GENE1296X	MCL1=myeloid cell differentiation protein
GENE1719X	TTG-2=Rhombotin-2=translocated in t(11;14)(p13;q11) T cell acute lymphocytic leukemia=cysteine rich protein with LIM motif
GENE1720X	TTG-2=Rhombotin-2=translocated in t(11;14)(p13;q11) T cell acute lymphocytic leukemia=cysteine rich protein with LIM motif
GENE3228X	JNK3=Stress-activated protein kinase
GENE3254X	Unknown UG Hs.145058 ESTs
GENE3255X	Unknown
GENE3256X	JAW1=lymphoid-restricted membrane protein
GENE3258X	JAW1=lymphoid-restricted membrane protein
GENE3259X	Unknown UG Hs.124922 ESTs
GENE3261X	Unknown
GENE3314X	Unknown
GENE3315X	FMR2=Fragile X mental retardation 2=putative transcription factor=LAF-4 and AF-4 homologue
GENE3318X	CD10=CALLA=Neprilysin=enkepalinase
GENE3325X	Unknown UG Hs.120245 Homo sapiens mRNA for KIAA1039 protein, partial cds
GENE3326X	Unknown UG Hs.105261 EST
GENE3327X	Unknown UG Hs.169565 ESTs, Moderately similar to !!!! ALU SUBFAMILY SB WARNING ENTRY !!!! [H.sapiens]
GENE3328X	Unknown UG Hs.136345 ESTs
GENE3329X	Unknown UG Hs.224323 ESTs, Moderately similar to alternatively spliced product using exon 13A [H.sapiens]
GENE3330X	Unknown
GENE3331X	Unknown UG Hs.208410 EST, Moderately similar to !!!! ALU SUBFAMILY SB WARNING ENTRY !!!! [H.sapiens]
GENE3332X	Unknown UG Hs.120716 ESTs
GENE3335X	myb-related gene A=A-myb
GENE3355X	Unknown
GENE3939X	Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene)
GENE3968X	Deoxycytidylate deaminase

VERY HIGH RELEVANCE

Fig. 2. Relevant genes to differentiate Activated B–like DLBCL from Germinal Centre B–like DLBCL (45 patients and 4026 attributes). The 50 most relevant genes for each feature selection method were selected. There is no order of relevance in the list.

DLBCL (GCL). Among these 45 examples, 22 belong to class GCL and 23 to class ACL.

The Relief algorithm, InfoGain and χ^2 methods have been applied to select the most relevant attributes to differentiate these sub–classes. These methods provided 25 common attributes, which will be considered as very high relevant, and they are enumerated in Figure 2.

Figures 1 and 2 do not include genes in common. It suggests that the genes required to differentiate among types of lymphoma cancer may be different to those distinguishing DLBCL subtypes. Nevertheless, several genes which experts had identified as having some functionality associated with lymphoma, are present in the subset, among them, TTG–2 and CD10. Furthermore, others like MCL1, JNK3 or FMR2, have not been linked to DLBCL.

3   Decision Trees

Decision trees are a useful technique in the context of supervised learning. They perform classification by a sequence of tests whose semantics are intuitively clear and easy to understand. Some tools, like J48, construct decision trees selecting the best attribute by using a statistical test to determine how well it alone classifies the training examples. Our experiments were performed by using the WEKA library for machine learning [18].

To avoid over–estimating the prediction accuracy that occurs when a model is trained and evaluated with the same samples, the "leave–one–out" testing method has been used. In this case 96–fold cross–validation and 45–fold cross–validation procedures are implemented when the lymphoma types and DLBCL subtypes are analysed respectively.

```
GENE1602X <= -0.44                  GENE646X <= -0.61                   GENE1622X <= -1.17
| GENE2426X <= 0.59                 | GENE844X <= -1.02: CLL            | GENE753X <= 0.61
| | GENE3959X <= 0.33: FL           | GENE844X > -1.02                  | | GENE753X <= -0.88
| | GENE3959X > 0.33: GCB           | | GENE2387X <= -0.68: RAT         | | | GENE653X <= -2.24: CLL
| GENE2426X > 0.59: CLL             | | GENE2387X > -0.68               | | | GENE653X > -2.24
GENE1602X > -0.44                   | | | GENE2374X <= 0.61: FL         | | | | GENE655X <= -1.34: RBB
| GENE563X <= -0.52                 | | | GENE2374X > 0.61: RBB         | | | | GENE655X > -1.34: CLL
| | GENE3701X <= -1.2: RAT          GENE646X > -0.61                    | | GENE753X > -0.88: FL
| | GENE3701X > -1.2: RBB           | GENE717X <= -0.5                  | GENE753X > 0.61
| GENE563X > -0.52                  | | GENE3639X <= -0.28: NIL         | | GENE694X <= 0.6: RAT
| | GENE717X <= -0.5: ABB           | | GENE3639X > -0.28: ABB          | | GENE694X > 0.6
| | GENE717X > -0.5                 | GENE717X > -0.5                   | | | GENE2403X <= 0.06: DLBCL
| | | GENE694X <= 1.54: DLBCL       | | GENE694X <= 1.54                | | | GENE2403X > 0.06: GCB
| | | GENE694X > 1.54: TCL          | | | GENE2402X <= 1.45: DLBCL      GENE1622X > -1.17
                                    | | | GENE2402X > 1.45: FL          | GENE694X <= -0.83: ABB
Tree Size = 15                      | | GENE694X > 1.54: TCL            | GENE694X > -0.83
Number of genes = 7 (from 4026)                                        | | GENE694X <= 1.54
Training Error = 2.08%              Tree Size = 17                      | | | GENE1610X <= -0.79: ABB
96-Fold CV Error = 20.84%           Number of genes = 8 (from 33)       | | | GENE1610X > -0.79: DLBCL
                                    Training Error = 5.38%              | | GENE694X > 1.54: TCL
                                    96-Fold CV Error = 26.05%
                                                                       Tree Size = 21
                                                                       Number of genes = 7 (from 8)
                                                                       Training Error = 6.79%
                                                                       96-Fold CV Error = 19.80%

        (a)                                  (b)                                 (c)
```

Fig. 3. Decision trees to differentiate lymphoma classes. All of the decision trees were generated using 96 patients. In (a) 4026 genes were used (the whole set); in (b) 33 genes (very high and high relevant); and in (c) only 8 genes (very high relevant). Labels were assigned to each type of lymphoma cancer, based on the study by Alizadeh et al.: Diffuse large B–cell lymphoma (DLBCL), Germinal centre B (GCB), NI. lymphoma node/tonsil (NIL), Activated blood B(ABB), Resting/activated T (RAT), Transformed cell lines (TCL), Follicular lymphoma (FL), Resting blood B (RBB) and Chronic lymphocytic leukaemia (CLL).

3.1   Prediction of Lymphoid Malignancies

Three decision trees were generated to differentiate among types of lymphoma. The first decision tree algorithm (Figure 3a) used the complete set of genes as input. However, the resulting tree comprises only 7 genes, producing an error rate of 2.08% (training set as test set), and 20.84% (leave–one–out method). The second decision tree (Figure 3b) selected 8 genes among the 33 genes extracted by the feature selection methods, making an error rate of 5.28% (training) and 26.05% (leave–one–out). The third decision tree (Figure 3c) provides an error rate of 6.79% (training) and 19.80% (leave–one–out). Thus, from a prediction point of view, only those genes categorized as very higly relevant allow the generation of the best decision trees.

The gene GENE694X (cyclin A) seems to be decisive in the prediction of lymphoma types, as it is the only one that appears in all of the decision trees. In the first one, differentiates DLBCL from TCL; in the second one, DLBCL and FL from TCL; and in the third one, RAT from DLBCL and GCB, and ABB and TCL from DLBCL (although the gen GENE1610X plays an important role to separate ABB from DLBCL). This machine learning method has remarkably recognised a key gene, which has been previously linked to the process of cell proliferation. Furthermore, a high protein expression of cyclin A has been associated with prognosis outcomes in non-Hodgkin's lymphomas [19].
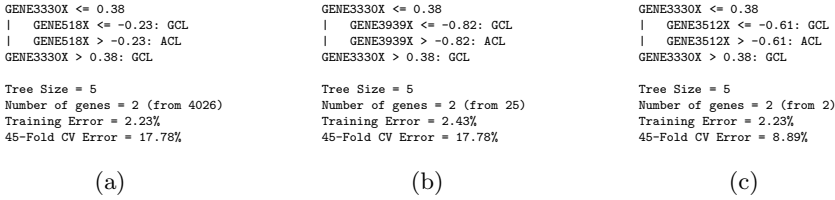
```
GENE3330X <= 0.38                  GENE3330X <= 0.38                  GENE3330X <= 0.38
|   GENE518X <= -0.23: GCL         |   GENE3939X <= -0.82: GCL        |   GENE3512X <= -0.61: GCL
|   GENE518X > -0.23: ACL          |   GENE3939X > -0.82: ACL         |   GENE3512X > -0.61: ACL
GENE3330X > 0.38: GCL              GENE3330X > 0.38: GCL              GENE3330X > 0.38: GCL

Tree Size = 5                      Tree Size = 5                      Tree Size = 5
Number of genes = 2 (from 4026)    Number of genes = 2 (from 25)      Number of genes = 2 (from 2)
Training Error = 2.23%             Training Error = 2.43%             Training Error = 2.23%
45-Fold CV Error = 17.78%          45-Fold CV Error = 17.78%          45-Fold CV Error = 8.89%
```

(a)   (b)   (c)

Fig. 4. Decision trees for DLBCL sub-classes. All of the decision trees were generated using 45 patients. In (a) 4026 genes were used (the whole set); in (b) 25 genes (very high relevant); and in (c) only 2 genes (1 was randomly chosen from the data).

3.2   Prediction of DLBCL Subtypes

Figure 4 illustrates three decision trees generated to differentiate among the DLBCL subtypes, ACL and GCL. Two genes were sufficient to build the trees. The gene GENE3330X has been included in all of the trees, which indicates its relevance to achieve this classification.

However, this gene can be combined with many others without considerably increasing the error rate. In fact, we randomly selected an a priori non–relevant gene, the gene GENE3512X, and the error rate was even lower than earlier, about 8.8% (leave–one–out). Therefore, one may state that the difference among the more important genes in terms of their relevance is very slight.

Based on medical research about the significance of specific genes to differentiate subtypes of DLBCL, 5 genes have been selected. They are those encoding CD10 (GENE3317X, GENE3318X and GENE3319X), BCL–6 (GENE3340X and GENE3341X), TTG–2 (GENE1719X and GENE1720X), IRF–4 (GENE1212X and GENE1213X) and BCL–2 (GENE2514X, GENE2536X, GENE2537X, GENE2538X), and some genes belonging to the BCL–2 family (GENE385X, GENE386X, GENE387X, GENE3619X and GENE3620X). The importance of these genes has been demonstrated by Azuaje by means of the simplified fuzzy ARTMAP model, which is a neural network–based model [3].

Results provided by the decision tree used only four genes (GENE1719X, GEN3318X, GENE3340X and GENE385X) and the error rate was 0% (training) and 26.67% (leave–one–out). The overfitting was very high, and therefore, these genes are not appropriate to predict accurately the subtype of DLBCL by using a decision tree.

4   Conclusions

A broad study of the database generated by Alizadeh et al. [2] was presented in this paper. It focused on both the feature selection and classification tasks.

From a biomedical point of view, the relevance of specific genes reported by Alizadeh et al. is not observed in our results. This is perhaps because other genes may also play an important role in processes associated with this disease. However, this conclusion may not be strongly supported by results, as these have

been obtained from a small amount of patients, in comparison to the number of genes.

These analyses indicate that the data are insufficient to state indisputable conclusions. Many subsets of genes can achieve a good prediction performance, although most of them would provide an overfitted decision tree. From a classification point of view, some genes are indeed very important, but more data should be included to support these observations. Alizadeh et al. inferred that a subset of genes could accurately differentiate among two subtypes of DLBCL. Nevertheless, none of such subsets have been identified by our KDD framework. The results also suggest that several subsets can attain the same classification aims. In fact, many decision trees can be built by using non–identified–as–relevant genes, producing similar error rates for the classification task. Furthermore, this research indicates that a deep study on the non–linear inter–relationship among genes might reveal interesting properties, as it has been discussed in [3]. With regard to the analysis of non–linear inter–relationships among genes for distinguishing lymphoma subtypes, we have recently built a neural network classifier based on 25 genes selected by the Relief method (with 3 nearest neighbours). This model, which consisted of two hidden layers and was tested with 45–fold cross–validation, produced an error rate equal to 0%.

This study highlights the importance of data mining techniques to extract interesting patterns from biological data, the significance of the results in contrast to statistics, and their future projection.

Acknowledgements

References

1. A. A. Alizadeh, M. Eisen, D. Botstain, P. O. Brown, and L. M. Staudt, "Probing lymphocyte biology by genomic-scale gene expression analysis," *Journal of clinical immunology*, no. 18, pp. 373–379, 1998.
2. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, truc Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. H. Jr., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, W. Dennis D, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large b–cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 2000.
3. F. Azuaje, "A computational neural approach to support discovery of gene function and classes of cancer," *IEEE Transactions on biomedical engineering*, vol. 48, no. 3, pp. 332–339, 2001.
4. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Belmont, CA: Wadsworth International Group, 1984.
5. A. D. Gordon, *Classification*. Chapman & Hall/CRC, 1999.

6. K. C. Gowda and G. Krishna, "Agglomerative clustering using the concept of mutual nearest neighborhood," *Pattern Recognition*, vol. 10, pp. 105–112, 1977.

7. M. A. Hall, "Correlation–based feature selection for machine learning," Ph.D., Department of Computer Science, University of Waikato, New Zealand, 1998.

8. J. Han and M. Kamber, *Data Mining – Concepts and Techniques*. Morgan Kaufmann, 2001.

9. N. L. Harris, E. S. Jaffe, J. Diebold, G. Flandrin, H. K. Muller-Hermelink, J. Vardiman, T. A. Lister, and C. D. Bloomfield, "World health organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: Report of the clinical advisory committee meeting–airlie house, virginia, november 1997," *Journal of clinical oncology*, vol. 17, pp. 3835–3849, 1999.

10. C. W. Hooper, R. C. Holman, M. J. Clarke, and T. L. Chorba, "Trends in non–hodgkin's lymphoma (NHL) and HIV–associated NHL deaths in the united states," *American Journal of Hematology*, vol. 66, pp. 159–166, 2001.

11. K. Kira and L. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Conference on Machine Learning*, 1992, pp. 249–256.

12. I. Kononenko, "Estimating attributes: analysis and extensions of relief," in *Proceedings of European Conference on Machine Learning*. Springer-Verlag, 1994.

13. F. Levi, F. Lucchini, E. Negri, and C. L. Vecchia, "Trends in mortaligy from non–hodgkin's lymphomas," *Leukemia Research*, vol. 26, pp. 903–908, 2002.

14. H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in *Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence*, 1995.

15. R. L. Rivest, "Learning decision lists," *Machine Learning*, vol. 1, no. 2, pp. 229–246, 1987.

16. M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutor, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub, "Diffuse large b–cell lymphoma outcome prediction by gene–expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68–74, 2002.

17. The Non-Hodgkin's Lymphoma Classification Project, "A clinical evaluation of the international of the international lymphoma study group. classification of non-hodgkin's lymphoma," *Blood*, vol. 89, no. 11, pp. 3909–3918, 1997.

18. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 2000.

19. D. Wolowiec, F. Bergera, P. Ffrench, P. Byron, and M. Ffrench, "CDK1 and cyclin A expression is linked to cell proliferation and associated with prognosis in non–hodgkin's lymphomas," *Leuk Lymphoma*, vol. 1–2, pp. 147–157, 1999.