On Feature Selection through Clustering

Richard Butterworth
Univ. of Massachusetts Boston,
Dept. of Computer Science,
Boston, MA 02125,
rickb@cs.umb.edu

Gregory Piatetsky-Shapiro
KDnuggets,
gregory@kdnuggets.com

Dan A. Simovici
Univ. of Massachusetts Boston,
Dept. of Computer Science,
Boston, MA 02125,
dsim@cs.umb.edu

Abstract

We study an algorithm for feature selection that clusters attributes using a special metric and then makes use of the dendrogram of the resulting cluster hierarchy to choose the most relevant attributes. The main interest of our technique resides in the improved understanding of the structure of the analized data and of the relative importance of the attributes for the selection process.

1 Introduction

The performance, robustness, and usefulness of classification algorithms are improved when relatively few features are involved in the classification. Thus, selecting relevant features for the construction of classifiers has received a great deal of attention. A lucid taxonomy of algorithms for feature selection was discussed in [17]; a more recent reference is [7]. Several approaches to feature selection have been explored, including wrapper techniques [12], support vector machines [6], neural networks [11] and prototype-based feature selection [9] that is close to our own approach.

The central idea of this work is to introduce an algorithm for feature selection that clusters attributes using a special metric and, then uses a hierarchical clustering for feature selection.

Hierarchical algorithms generate clusters that are placed in a cluster tree, which is commonly known as a *dendrogram*. Clusterings are obtained by extracting those clusters that are situated at a given height in this tree.

Our intent is to show that good classifiers can be built by using a small number of attributes located at the centers of the clusters identified in the dendrogram. This type of data compression can be achieved with little or no penalty in terms of the accuracy of the classifier produced. The clustering of attributes helps the user to understand the structure of data, the relative importance of attributes. Alternative feature selection method mentioned above are excellent in reducing the data without having a severe impact on the accuracy of classifiers; however, such methods cannot identify how attributes are related to each other.

An *object system* is a pair $\mathcal{S} = (S, H)$, where S is set called the set of objects of $\mathcal{S}$, $H = \{A_1, \ldots, A_m\}$ is a set of mappings defined on S. We assume that for each mapping A_i (referred to as an attribute (or a feature) of $\mathcal{S}$) there exists a nonempty set E_i called the domain of A_i such that $A_i : S \longrightarrow E_i$ for $1 \leq i \leq m$. The value of an attribute A_i on an object t is denoted by $t[A_i]$. This terminology is consistent with the terminology used in relational databases, where a table can be regarded as an object system; however, the notion of object system is more general because objects have an identity as members of the set S, instead of being regarded as just m-tuples of values. In this spirit, we shall refer to $t[A_i]$ as *projection of* t *on* A_i.

Let S be a set. A *partition on* S is a non-empty collection of subsets of S indexed by a set I, $\pi = \{B_i \mid i \in I\}$ such that $\bigcup_{i \in I} B_i = S$ and $i \neq j$ implies $B_i \cap B_j = \emptyset$. The sets B_i are commonly referred to as the *blocks of the partition* π. The set of partitions on S is denoted by $\mathsf{PART}(S)$.

An attribute A of an object system $\mathcal{S} = (S, H)$ generates a partition π^A of the set of objects S, where two objects belong to the same block of π^A if they have the same projection on A. We denote by B_a^A the block of π^A that consists of all tuples of S whose A-component is a. Note that for

relational databases, π^A is the partition of the set of rows of a table that is obtained by using the **group by** A option of **select** in standard SQL.

The set of partitions of a set can be naturally equipped with a partial order. For $\pi, \sigma \in \mathsf{PART}(S)$ we write $\pi \leq \sigma$ if every block B of π is included in a block of σ, or equivalently, if every block of σ is an exact union of blocks of π. This partial order generates a lattice structure on $\mathsf{PART}(S)$; this means that for every two partitions $\pi, \pi' \in \mathsf{PART}(S)$ there is a least partition π_1 such that $\pi \leq \pi_1$ and $\pi' \leq \pi_1$ and there is a largest partition π_2 such that $\pi_2 \leq \pi$ and $\pi_2 \leq \pi'$. The first partition is denoted by $\pi \vee \pi'$, while the second is denoted by $\pi \wedge \pi'$.

2 Distance between partitions and the Pearson index

To introduce a metric on the set of partitions of a finite set we define the mapping $v : \mathsf{PART}(S) \longrightarrow \mathbb{R}$ by $v(\pi) = \sum_{i=1}^{n} |B_i|^2$, where $\pi = \{B_1, \ldots, B_n\}$. The mapping v is a lower valuation on $\mathsf{PART}(S)$, that is,

$$v(\pi \vee \sigma) + v(\pi \wedge \sigma) \geq v(\pi) + v(\sigma) \tag{1}$$

for $\pi, \sigma \in \mathsf{PART}(S)$.

For every lower valuation v the mapping $d : (\mathsf{PART}(S))^2 \longrightarrow \mathbb{R}$ defined by $d(\pi, \sigma) = v(\pi) + v(\sigma) - 2 \cdot v(\pi \wedge \sigma)$ is a metric on $\mathsf{PART}(S)$ (see [3, 2, 14]). We will refer to d as the *Barthélemy-Montjardet distance*.

Using the cardinalities of the blocks of the partitions we can write

$$d(\pi, \sigma) = \sum_i |B_i|^2 + \sum_j |C_j|^2 - 2 \sum_i \sum_j |B_i \cap C_j|^2,$$

where $\pi = \{B_1, \ldots, B_n\}$ and $\sigma = \{C_1, \ldots, C_p\}$. This metric was used for the development of an incremental clustering algorithm[15]. In this paper we use it to cluster attributes.

For a partition $\pi = \{B_1, \ldots, B_n\}$ denote by M_π and m_π the largest and the smallest size of a block of π.

Let $\pi = \{B_1, \ldots, B_n\}$, $\sigma = \{C_1, \ldots, C_p\}$ be two partitions. The *contingency matrix* of π, σ is the matrix $P_{\pi,\sigma}$ whose entries are given by $p_{ij} = |B_i \cap C_j|$ for $1 \leq i \leq n$ and $1 \leq j \leq p$. The Pearson χ^2 association index can be written in our framework as:

$$\chi^2_{\pi,\sigma} = \sum_i \sum_j \frac{(p_{ij} - |B_i||C_j|)^2}{|B_i| \cdot |C_j|}.$$

It is well-known (See [1]) that the asymptotic distribution of this index is a χ^2-distribution with $(n-1)(p-1)$ degrees of freedom.

Theorem 2.1 *Let S be a finite set and let $\pi, \sigma \in \mathsf{PART}(S)$, where $\pi = \{B_1, \ldots, B_n\}$ and $\sigma = \{C_1, \ldots, C_p\}$. We have:*

$$\frac{v(\pi)+v(\sigma)-d(\pi,\sigma)}{2M_\pi M_\sigma} - 2np + |S|^2 \leq \chi^2_{\pi,\sigma} \leq \frac{v(\pi)+v(\sigma)-d(\pi,\sigma)}{2m_\pi m_\sigma} - 2np + |S|^2.$$

Proof. Note that $\chi^2_{\pi,\sigma} = \sum_i \sum_j \frac{p_{ij}^2}{|B_i| \cdot |C_j|} - 2np + |S|^2$. Since $m_\pi m_\sigma \leq |B_i||C_j| \leq M_\pi M_\sigma$, we have:

$$\frac{p_{ij}^2}{M_\pi M_\sigma} \leq \frac{p_{ij}^2}{|B_i| \cdot |C_j|} \leq \frac{p_{ij}^2}{m_\pi m_\sigma}.$$

Thus,

$$\frac{v(\pi \wedge \sigma)}{M_\pi M_\sigma} - 2np + |S|^2 \leq \chi^2_{\pi,\sigma} \leq \frac{v(\pi \wedge \sigma)}{m_\pi m_\sigma} - 2np + |S|^2$$

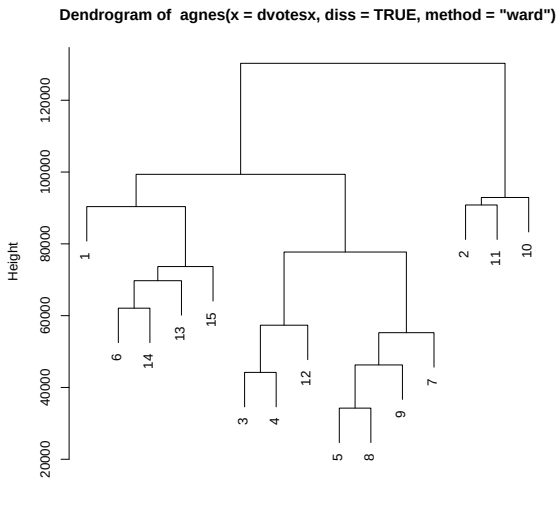
Since $d(\pi, \sigma) = v(\pi) + v(\sigma) - 2 \sum_i \sum_j p_{ij}^2$, the desired equality follows immediately.

The Pearson coefficient decreases with the distance and, thus, the probability that π and σ are independent increases with the distance. This suggest that partitions that are correlated are close in the sense of the Barthélemy-Montjardet distance; therefore, if attributes are clustered using the corresponding distance between partitions we could replace clusters with their centroids and, thereby, drastically reduce the number of attributes involved in a classification without significant decreases in accuracy of the resulting classifiers.

3 Experimental Validation

We experimented with several data sets from the UCI dataset repository [4] and, due to space limitations we discuss only the results obtained with the votes and zoo datasets, which have a relative small number of categorical features. In each case, starting from the matrix $(d(\pi^{A_i}, \pi^{A_j}))$ of Barthélemy-Montjardet distances between the partitions of the attributes $A_1, \ldots, A_n$, we clustered the attributes using AGNES, an agglomerative hierarchical algorithm [10] implemented as a component of the cluster package of system R (see [13]).

Clusterings were extracted from the tree produced by the algorithm by cutting the tree at various heights starting with the maximum height of the tree created above (corresponding to a single cluster) and working down to a height of 0 (which consists of single-attribute clusters). A 'representative' attribute was created for each cluster as the attribute that has the minimum total distance to the other members of the cluster, again using the Barthélemy-Montjardet distance. The J48 and the Naïve Bayes algorithms of the WEKA package [16] were used for constructing classifiers on data sets obtained by projecting the initial data sets on the sets of representative attributes.



1	handicapped_infants
2	water_project_cost_sharing
3	budget_resolution
4	physician_fee_freeze
5	el_salvador_aid
6	religious_groups_in_schools
7	anti_satellite_test_ban
8	aid_to_nicaraguan_contras
9	mx_missile
10	immigration
11	synfuels_corporation_cutback
12	education_spending
13	superfund_right_to_sue
14	crime
15	duty_free_exports

Figure 1. Dendrogram of votes **Dataset using AGNES and the Ward method**

The dataset votes records the votes of 435 US Congressmen on 15 key questions, where each attribute can have the value ”y”,”n”, or ”?” (for abstention), and each Congressman is classified as a democrat or republican. It is interesting to note that by applying the AGNES clustering algorithm with the Ward method of computing the inter-cluster distance the voting issues group naturally into clusters that involve larger issues, as shown in Figure 1. For example, ”el_salvador_aid”,”aid_to_nicaraguan_contras”, ”mx_missile”, and ”anti_satellite_test_ban” are grouped quite early into a cluster that can be described as dealing with defence policies. Similarly, social budgetary legislation issues such as ”budget_resolution”, ”physician_fee_freeze”, and ”education_spending”, are grouped together.

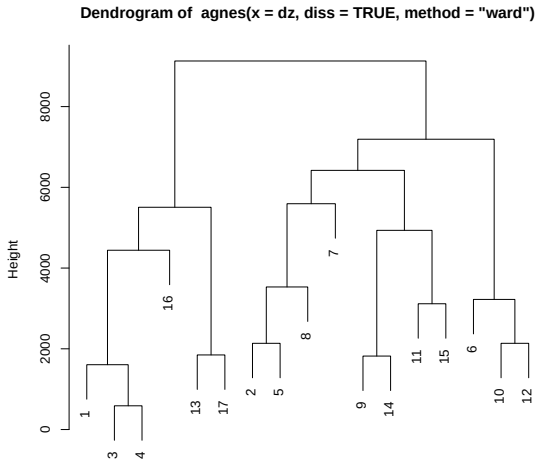
Two types of classifiers (J48 and Naïve Bayes) were generated using ten-fold cross validation by extracting centrally located attributes from cluster obtained by cutting the dendrogram at successive levels. The accuracy of these classifiers is shown in Table 1. This experiment shows that our method identifies the most influential attribute 5 (in this case ”el_salvador_aid”). So, in addition to reducing number of attributes, the proposed methodology allows us to assess the relative importance of attributes.

Attribute Set (class attribute not listed)	Classifiers	
	J48%	NB%
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15	96.78	90.34
1,2,3,4,5,6,7,9,10,11,12,13,14,15	96.78	91.03
1,2,3,4,5,6,7,10,11,12,13,14,15	96.55	91.26
1,2,4,5,6,7,10,11,12,13,14,15	95.17	92.18
1,2,4,5,6,10,11,12,13,14,15	95.17	92.64
1,2,4,5,6,10,11,13,14,15	95.40	92.18
1,2,6,8,10,11,13,14,15	86.20	85.28
1,2,8,10,11,13,14,15	86.20	85.74
1,2,8,10,11,14,15	84.13	85.74
1,2,8,10,11,14	83.69	85.74
2,8,10,11,14	83.67	84.36
2,5,10,11	88.73	88.50
2,5,10	84.82	84.82
2,5	84.82	84.82
5	84.82	84.82

Table 1. Accuracy of classifiers for the Votes **dataset constructed on attribute sets obtained by clustering**

A similar study was undertaken for the zoo database, after eliminating the attribute animal which determines uniquely the type of the animal. Starting from a dendrogram build by using the Ward method shown in Figure 2 we constructed J48 and Naïve Bayes classifiers for several sets of attributes obtained as successive sections of the cluster tree. The results are shown in Table 2. Note that attributes that are biologically correlated (e.g. hair,milk, and eggs, or aquatic (6), breathes (10), and fins(12)) belong to relatively early clusters).

We believe that the main interest of the proposed approach to attribute selection is the possibility of the supervision of the process allowing the user to opt between quasi-equivalent attributes (that is, attributes that are close relatively to the Barthélemy-Montjardet metric) in order to produce more meaningful classifiers. We compared our approach with two existing attribute set selection techniques: the correlation-based feature (CSF) selection (developed in [8] and incorporated in the WEKA package and the wrapper technique, using the “best-first” and the greedy method as search methods, and the J48 classifier for the classifier incorporated by the wrapper. For the zoo data set we obtained identical attribute sets with either “best-first” or with



		9	backbone
1	hair	10	breathes
2	feathers	11	venomous
3	eggs	12	fins
4	milk	13	legs
5	airborne	14	tail
6	aquatic	15	domestic
7	predator	16	catsize
8	toothed	17	type

Figure 2. Dendrogram of zoo dataset using AGNES and the Ward method

Attribute Set (class attribute not listed)	Classifiers	
	J48%	NB%
1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16	92.07	93.06
1,2,4,5,6,7,8,9,10,11,12,13,14,15,16	92.07	92.07
2,4,5,6,7,8,9,10,11,12,13,14,15,16	87.12	88.11
2,4,5,6,7,8,9,10,11,12,13,15,16	87.12	88.11
2,4,6,7,8,9,10,11,12,13,15,16	88.11	87.12
2,4,6,7,8,9,10,11,13,15,16	91.08	91.08
2,4,6,7,8,9,10,11,13,16	89.10	90.09
2,4,7,8,9,10,11,13,16	86.13	90.09
2,4,7,9,10,11,13,16	84.15	90.09
2,4,7,9,10,11,13	87.12	89.10
4,5,7,9,10,11	88.11	88.11
4,5,7,9,10	88.11	90.09
4,5,9,10	89.10	91.09
4,5,10	73.26	73.26
4,10	73.26	73.26
4	60.39	60.39

Table 2. Accuracy of classifiers for the zoo dataset constructed on attribute sets obtained by clustering

Attribute Selection	Experimental Results
CSF	Attr. set: 1,2,4,8,9,10,12,13,14 Accuracy for J48: 91.08% Accuracy for NB: 95.04%
Wrapper with J48	Attr. set: 1,2,4,8,9,12,13 Accuracy for J48: 96.03% Accuracy for NB: 92.07%

Table 3. Accuracy of classifiers obtained through attribute selection techniques

the greedy method. The results are shown in Table 3.

These results suggest that this method is not as good for accuracy as the the wrapper method or CSF. However, the tree of attributes helps to understand the relationships between attributes and their relative importance.

4 Conclusion and Future Work

Attribute clustering help to build classifiers in a semi-supervised manner allowing analysts a certain degree of choice in the selection of the features that may be considered by classifiers, and illuminating relationships between attributes and their relative importance for classification.

As stated in [7], in early studies of relevance published in the late 90s [5, 12], few applications explored data with more than 40 attributes. With the increased interest of data miners in bio-computing in general, and in microarray data in particular, classification problems that involve thousands of features and relatively few examples came to the fore. We intend to apply our techniques to this type of data.

References

[1] A. Agresti. *An Introduction to Categorical Data Analysis*. John Wiley, New York, 1997.

[2] J. Barthélemy. Remarques sur les propriétés metriques des ensembles ordonnés. *Math. Sci. hum.*, 61:39–60, 1978.

[3] J. Barthélemy and B. Leclerc. The median procedure for partitions. In *Partitioning Data Sets*, pages 3–34, Providence, 1995. American Mathematical Society.

[4] C. L. Blake and C. J. Merz. *UCI Repository of machine learning databases*. University of California, Irvine, Dept. of Information and Computer Sciences, http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998.

[5] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, pages 245–271, 1997.

[6] M. Brown, W. Grundy, D. Lin, N. Cristiani, C. W. Sugnet, T. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*, 97:262–267, 2000.

[7] E. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J. of Machine Learning Research*, pages 1157–1182, 2003.

[8] M. A. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, Hamilton, New Zealand, 1999.

[9] B. Hanczar, M. Courtine, A. Benis, C. Hannegar, K. Clement, and J. Zucker. Improving classification of microarray data using prototype-based feature selection. *SIGKDD Explorations*, pages 23–28, 2003.

[10] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data – An Introduction to Cluster Analysis*. Wiley Interscience, New York, 1990.

[11] J. Khan, J. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westerman, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine*, 7:673–679, 2001.

[12] R. Kohavi and G. John. Wrappers for feature selection. *Artificial Intelligence*, pages 273–324, 1997.

[13] J. Maindonald and J. Brown. *Data Analysis and Graphics Using R*. Cambridge University Press, Cambridge, 2003.

[14] B. Monjardet. Metrics on parially ordered sets – a survey. *Discrete Mathematics*, 35:173–184, 1981.

[15] D. Simovici and N. Singla. Metric incremental clustering of categorical data. In *Proceedings of ICDM*, pages 523–527, 2004.

[16] I. H. Witten and E. Frank. *Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, 2000.

[17] D. Zongker and A. Jain. Algorithms for feature selection: An evaluation. In *Proceedings of the International Conference on Pattern Recognition*, pages 18–22, 1996.