UDC: 007.52
Original scientific paper

EMPIRICAL EVALUATION OF CLUSTERING ALGORITHMS*

Andreas Rauber, Elias Pampalk

Vienna University of Technology, Department of Software Technology
E-mail: {andi, elias}@ifs.tuwien.ac.at

Jan Paralic

Technical University of Kosice, Department of Cybernetics and Artificial Intelligence
E-mail: paralic@tuke.sk

Unsupervised data classification can be considered one of the most important initial steps in the process of data mining. Numerous algorithms have been developed and are being used in this context in a variety of application domains, albeit, only little evidence is available as to which algorithms should be used in which context, and which techniques offer promising results when being combined for a given task. In this paper we present an empirical evaluation of some prominent unsupervised data classification techniques with respect to their usability and the interpretability of their result representation.

Keywords: data mining, cluster analysis, hierarchical agglomerative clustering, Bayesian clustering, Self-Organizing Map (SOM), growing hierarchical SOM, generative topographic mapping.

1. INTRODUCTION

Detecting unknown patterns in large datasets is becoming an increasingly important factor with respect to the wider availability of large data collections and the need to extract knowledge hidden within them. With a wealth of highly sophisticated tools from areas such as statistics, fuzzy logics, expert systems, neural networks and other AI-related techniques available, unsupervised classification is frequently employed as one of the first steps in the iterative process of mining the data. The main goal of this step is to provide an overview of the characteristics of the data, its inherent structure, and clusters in order to get an at least intuitive feeling for a given dataset before proceeding with other methods of data analysis, if the main task is broader than identifying the structure of the dataset.

Even within the family of unsupervised classification techniques, a variety of methods offer themselves for this process. Most of these methods are well-founded and tested techniques that have been employed in numerous applications so far, with new variations of these techniques being developed constantly, each addressing specific shortcomings of their ancestors. The decision which method to use depends to a large degree on assumptions relating to the distribution of the data, some of which may

*) This work is supported by the Austrian Institute for East- and Southeast Europe as part of the CLUE Project.

be inherent in one or the other method, on processing requirements, or on their ability to handle non-numerical or incomplete data. Once these decisions have been made, and the number of techniques has been narrowed down to a few to choose from, the decision should be and usually will be based on the usability of the various methods, their quality of result representation and the ease (and tool-support) by which these results can be interpreted.

While the mathematical properties of the various approaches are usually well-analyzed [7], little work has so far been published with the focus on result representation and evaluation. In this paper we address this problem by comparing the result representation of a number of popular unsupervised classification techniques and their modifications. As mentioned above, the focus is not so much on the correctness of the results or the assumption inherent in the algorithms used, but rather their robustness in terms of parameter selection, their ease of handling and the information to be gained from their result representation. Where possible we try to abstract from the specific implementations used, and rather concentrate on the general characteristics of the approaches.

For this purpose we selected five methods, namely (1) a complete linkage Hierarchical Agglomerative Clustering; (2) Bayesian Clustering using AutoClass; (3) the Self-Organizing Map (SOM), a prominent unsupervised neural network model mapping high-dimensional data onto a two-dimensional plane; (4) the Growing Hierarchical SOM, a recent extension to the SOM model allowing hierarchical cluster analysis; and (5) Generative Topographic Mapping, a probabilistic model for generating topology preserving mappings.

The remainder of this paper is structured as follows. In Section 2 we present an overview of the methods selected for these experiments. This is followed by a presentation of experimental results in Section 3, where we use an artificial data set as a toy example to present the main characteristics of the algorithms. After this initial evaluation we use these algorithms to analyze a high-dimensional dataset from the field of text classification in Section 4. An analysis of the methods' result representation and their interpretability is presented in Section 5, followed by some conclusions and lessons learned in Section 6.

2. METHODS

In this section we present a brief introduction into the methods used and provide pointers to more detailed descriptions of the algorithms.

2.1. Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering builds a hierarchical classification of data items by a series of mergers (agglomerations).

The clusters most similar to each other are merged together to form one cluster, and this is repeated starting with clusters with only one vector up to one cluster containing all vectors. One possibility to calculate distances between the clusters is the complete-

linkage method [9], which is the maximum of all pair wise distances between vectors taken from the compared clusters. Another possibility, for example, is the single-linkage method [19], which, in contrast to the complete-linkage, uses the minimum distance. We use the complete-linkage algorithm for the further analysis in this paper.

Due to its age the Hierarchical Agglomerative Clustering algorithm and its derivatives have been widely employed and are included in a number of commercial data mining packages.

2.2. Bayesian Classification (AutoClass)

The program AutoClass [3], Automatic Class discovery from Data, uses Bayesian probability theory to provide an extensible approach to problems such as classification and general mixture separation.

AutoClass describes classes by probability distributions over the attributes of the data items. The calculation of the probability of each data item's membership to each class provides a more spacing classification than absolute partitioning techniques.

The user can choose from default models or specify a class probability distribution function by associating attribute sets with supplied likelihood function terms. AutoClass then searches in the space of class numbers and parameters for the maximally probable combination. It returns the set of class probability function parameters, and the class membership probabilities for each data instance.

AutoClass has been employed successfully in a number of applications [2].

2.3. Self-Organizing Map

The Self-Organizing Map (SOM) [10] is an unsupervised neural network mapping high dimensional input data, usually onto a two-dimensional output space, while preserving relations between the data items both. The cluster structure within the data as well as the inter-cluster similarity are visible from the resulting topology preserving mapping.

The SOM consists of units (neurons), which are arranged as a two-dimensional rectangular or hexagonal grid. During the training process vectors from the dataset are presented to the map in random order. The unit most similar to a chosen vector is selected as the winner and adopted to match the vector even better. Then units in the neighborhood of the winner are slightly adopted as well. The trained SOM provides a mapping of the data space onto a two-dimensional plain in such a way that similar data points are located close to each other. Additional visualization techniques such as the U-Matrix [20], Adaptive Coordinates [12], or cluster connections [13] aid the user in understanding the cluster structure. Furthermore methods like LabelSOM [15] allow the automatic extraction of cluster descriptions based on the attributes.

The SOM and its variants have been employed many times in a wide variety of domains, such as financial, medical or time series data analysis [4, 11, 18].

2.4. Growing Hierarchical Self-Organizing Map

The Growing Hierarchical Self-Organizing Map (GHSOM) [5] is a new variation of the SOM. The basic idea is to allow the SOM to grow in width and depth creating a flexible hierarchical structure where the size of the feature map is determined automatically.

The GHSOM grows in width by adding new units to the SOM during the training process in areas where they are needed, similar to the Growing Grid network [6]. It furthermore grows in depth by training a new GHSOM for units representing larger clusters, thus automatically detecting and mirroring the hierarchical structure inherent in the data. It thus combines the advantages of the Growing Grid and the Hierarchical Feature Map [14] while overcoming some of their limitations.

Being one of the most recent enhancements of the SOM method, the GHSOM has not been employed intensively, but it has shown to produce very promising results in high-dimensional data classification tasks [16].

2.5. Generative Topographic Mapping

The Generative Topographic Mapping (GTM) algorithm [1] consists of a constrained mixture of Gaussians in which the model parameters are determined by maximum likelihood using the Expectation Maximization (EM) algorithm. A set of points in the two-dimensional latent space, which are similar to the units of the SOM, are mapped by a non-linear and continuous function into the data space.

The result of the GTM algorithm is a density function for each data item in the two dimensional latent space. To simplify the visualization of the data, often only the means and modes of these distributions are used, obtaining visualization similar in spirit to the SOM.

In spite of being statistically well-founded, the GTM has not yet been applied to and evaluated on too many different real-world applications.

3. A TOY EXAMPLE: CLUSTERING ANIMALS

This artificial dataset consists of 16 animals described by 13 attributes such as size, number of legs etc. [17]. This dataset is easy to handle and to evaluate different parameter settings. Furthermore the results are intuitively interpretable, allowing us to straightforwardly understand and compare the basic functionality of the chosen methods.

3.1. Hierarchical Agglomerative Clustering

Hierarchical Agglomerative Clustering merges the animals' cluster at different levels of similarity, with the most similar animals being merged in the beginning, and the various clusters being merged hierarchically in subsequent steps.

The best results are achieved with complete linkage with raw data and the Euclidean distance. The two main branches, birds and mammals, are clearly separated.

As can be seen in Figure 3.1.1 the results are very easy to interpret. Classes merged on lower levels are more similar than ones merged on higher levels. Note how owl and *falcon*, as well as *horse* and *zebra* are merged on the lowest level, since identical vectors describe these animals.

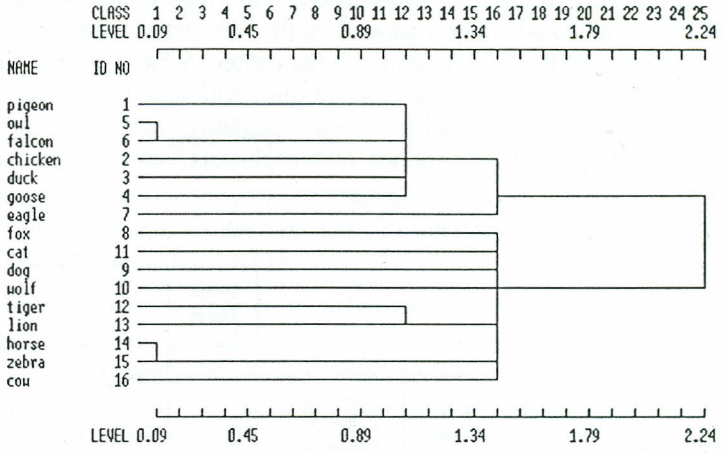

Figure 3.1.1: Dendrograms

3.2. Bayesian Classification (AutoClass)

AutoClass finds the best model with two classes, which represent the mammals and the birds. Figure 3.2.1 lists part of the description for the mammals' class generated by AutoClass. The attributes with the highest influence are *four legs*, *hair* and *no feathers*. The influence is the cross entropy or Kullback-Leibler distance between the class and full database probability distributions.

Attribute Name	Influence	Attribute Name	Influence
has four legs	0.395	can run	0.146
has hair	0.395	has mane	0.079
has no feathers	0.395	has hooves	0.056
cannot fly	0.229	cannot swim	0.048
is not small	0.216	is hunter	0.005

Figure 3.2.1: AutoClass Mammals Description

3.3. Self-Organizing Map

For result representation we chose the simple default SOM output, mapping the data points onto grid-like tables, with the cells representing the units of the map. De-

pending on the desired resolution of the map, sizes between 1x2 and 5x5 can be chosen.

Using a 1x2 map (cf. Figure 3.3.1) the dataset is split into two clusters of birds and mammals. Figure 3.3.2 shows the result using a 3x3 SOM, providing a finer distinction between clusters. We again find the mammals located in the upper half of the SOM to be separated from the birds in the lower half. Within the two big clusters of birds and mammals there are further sub-clusters, with, for example, the big mammals such as *cow*, *zebra* and *horse* being separated on the left part of the cluster whereas the smaller ones are more to the left and down to the center. A 5x5 map (cf. Figure 3.3.3) will result in every animal being on a separate unit, with again a very clear topology representation depicting the relationship between the data points. Using the LabelSOM method would create further descriptions of the clusters, similar to the results presented in [15].

Eagle Falcon Owl Goose Duck Chicken Pigeon	Cow Zebra Horse Lion Tiger Cat Wolf Dog Fox

Figure 3.3.1: 1x2

Wolf Dog	Lion Tiger	Cow Zebra Horse
Fox	Cat	
Eagle	Falcon Owl Pigeon	Goose Duck Chicken

Figure 3.3.2: 3x3

Zebra Horse		Tiger	Lion	Dog
	Cow		Wolf	
		Cat	Fox	
Duck	Goose			Eagle
Chicken	Pigeon		Falcon Owl	

Figure 3.3.3: 5x5

3.4. Growing hierarchical Self-Organizing Map

The GHSOM creates a hierarchical view of the dataset. On the first layer of the GHSOM, a rather rough separation of clusters is found, resulting in a 2x2 SOM with two units representing the birds and one unit representing the mammals (cf. Figure 3.4.1)

Whereas the unit representing *chicken*, *duck* and *goose* already provides a rather detailed representation of the data, the other units are expanded in a second layer, resulting in a more detailed representation for those sub-clusters.

For example, the second-layer SOM representing the mammals grows again to a size of again 2x2, separating the big mammals with *hooves* such as *horse*, *zebra* and *cow* from the *medium sized hunting* mammals (cf. Figure 3.4.2).

Pigeon Owl Falcon Eagle	Chicken Duck Goose
Fox Dog Wolf Cat Tiger Lion Horse Zebra Cow	

Figure 3.4.1: 1st Layer

Lion	Horse Zebra Cow
Fox Dog Wolf Tiger	Cat

Figure 3.4.2: 2nd Layer

3.5. Generative Topographic Mapping

GTM generates a probability density function for each pattern. In Figure 3.5.1 the shading represents the sum of the densities of each data item. The symbols are placed at the means of the single density functions.

As can be seen there is a big division between the mammals on the right side and the birds on the left side. Within these clusters there are further clusters. Most obvious is the separation between *medium sized hunting* birds such as *owl*, *falcon* and *eagle* from the smaller birds such as *pigeon*, *chicken*, *duck* and *goose*. But also within the mammals cluster some sub-clusters can be recognized. In the center are the smaller ones while the big hunters such as tiger and lion are located on the top. In the lower regions of the mammals cluster are the *big non-hunting* animals such as *zebra*, *cow* and *horse*. Note that because *zebra* and *horse* have identical vectors they only appear as one mark on the map. The same applies to *owl* and *falcon*.

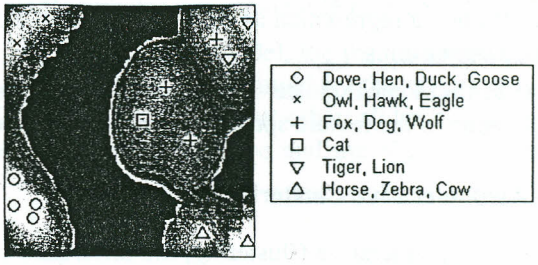

Figure 3.5.1: GTM Animals Map

3.6. Comparison

As can be expected, all methods have revealed the same basic clusters in the data. Figure 3.6.1 provides a projection of all results into the 5x5 SOM. Notice how the main separation between mammals and birds is found by all methods. Further subdivisions, although not completely identical, are found by the other methods, except for AutoClass, which separates the dataset only into two large sub clusters. *Zebra*, *horse* and *cow* are clustered together by all other methods. GTM and SOM order the *pigeon* to the *duck*, *goose* and *chicken* cluster while Hierarchical Agglomerative Clustering

with complete linkage and GHSOM assign it to the *eagle*, *falcon* and *owl*. Noticeable also is that only GHSOM considers *tigers* and *foxes* more similar than *foxes* and *cats* or *tigers* and *lions*.

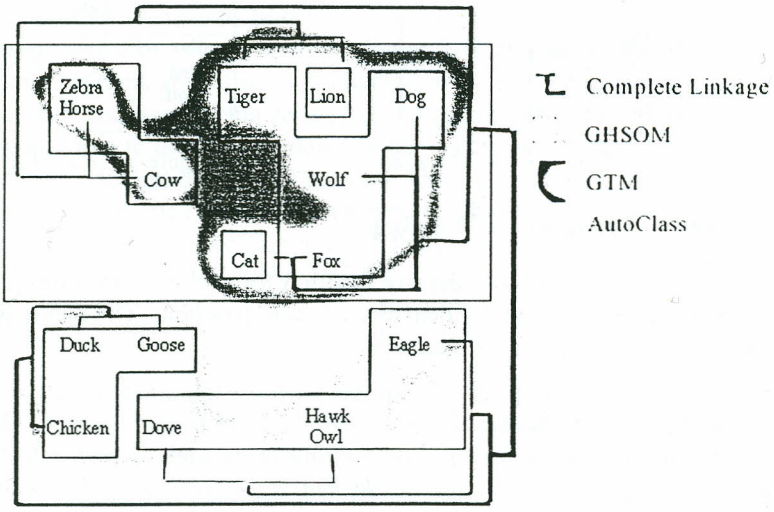

Figure 3.6.1: Comparison based on SOM 5x5

4. CLUSTERING HIGH-DIMENSIONAL DATA

For the second experimental setup we use a rather high-dimensional real-world dataset. It represents 420 newspaper articles of the TIME Magazine from the 1960's[1] with the articles being represented as word histograms resulting in a dataset of 420 vectors with 5923 dimensions.

The basic goal is to find topical clusters, i.e. articles covering similar topics as expressed by similar location in the feature space spanned by the words

4.1. Hierarchical Agglomerative Clustering

Using Hierarchical Agglomerative Clustering, the difference between normalizing the data vectors to unit length 1 and non-normalized data becomes obvious on first glance (compare Figures 4.1.1 and 4.1.2). The reason for this is that without normalization the length of the documents has a big influence in the revealed structure. For the remainder of the experiments presented in this section we will use normalized data.

The magnified area in Figure 4.1.2 shows a sub-tree. The articles in this sub-tree are identical to ones mapped onto the units on the first and second row in the eighth column in Figure 4.3.1. As a whole the clusters are very similar to those generated by the SOM.

[1] Available at http://www.ifs.tuwien.ac.at/~andi/somlib/.

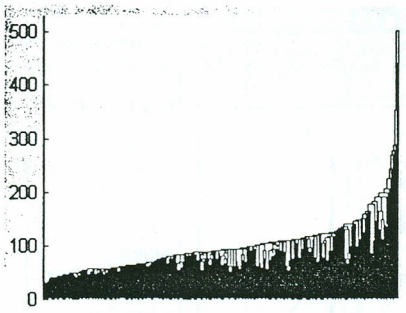

Figure 4.1.1: Non Normalized Data

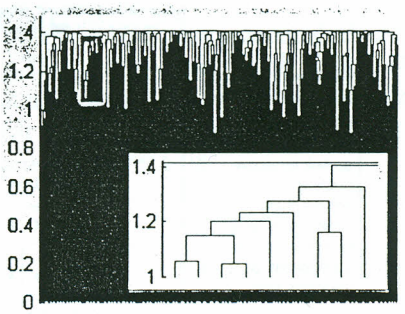

Figure 4.1.2: Normalized Data

4.2. Bayesian Classification (AutoClass)

AutoClass searches for the best model in the model space defined by the user. One major parameter for this model space is the distributions of the attributes. AutoClass offers predefined choices mainly for Gaussian and related distributions. Unfortunately, the distribution of the attributes in the dataset is much better described as exponential. Looking at one attribute, for example Vietnam, we find most articles do not contain this word, so their corresponding values are zero. Only a few from the 420 articles actually have a value bigger than zero.

Other than problems modeling the data there are also some problems handling the input and output. AutoClass seems to not have been developed for such high dimensional data. The report files generated, containing the found information in ASCII format, reach sizes up to 30MB, making it necessary to develop tools to handle them.

Also the input is not so easy to handle, since it requires a description of each attribute. With almost 6000 attributes it is impossible to generate the input manually.

Considering the problems with the model, the results are quite good. Some of the found clusters were very similar to those found by the other methods, for example a cluster about NATO and the Cold War. But others did not make sense; for example, documents about the monarchy in Morocco and the Vietnam War were in the same cluster.

4.3. Self-Organizing Map

The 10x15 map (cf. Figure 4.3.1) provides a good and intuitive overview of the TIME data. The main topical clusters can be easily detected, especially with the additional help provided by the LabelSOM technique.

We find, that the SOM has succeeded in creating a topology-preserving mapping of the document collection, i.e. we find documents on similar topics located on the same or neighboring units. For example, all articles mapped onto the units in the lower left corner of the map deal with problems in South Vietnam, with some units representing articles on the Vietnam War and other units covering the government crackdown on Buddhist monks.

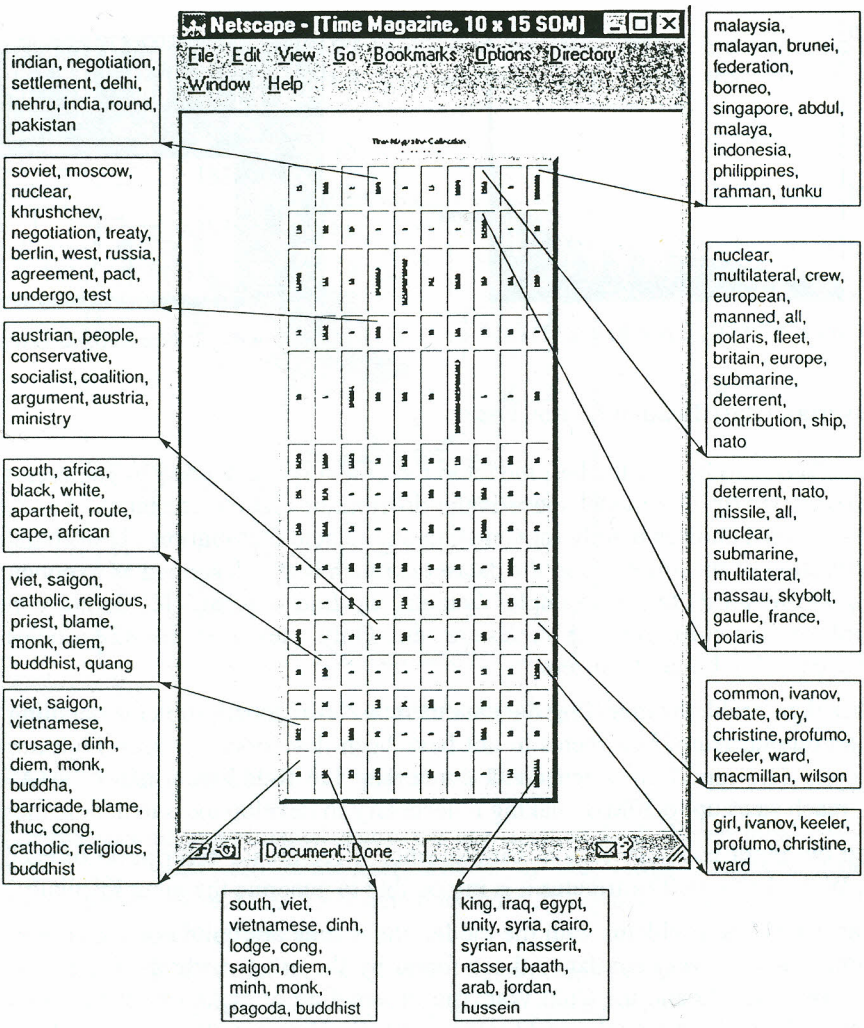

Figure 4.3.1: 10x15 SOM

As another example, consider the articles on the unit in the first row and the fourth, which all deal with the relationship between India and Pakistan and the Kashmir conflict. Several further topical clusters can be identified on the map, such as European Politics, the relationship between the east and the west during the Cold War, or the situation in the Middle East.

4.4. Growing hierarchical Self-Organizing Map

The top-level map (cf. Figure 4.4.1) evolved to a 2x5 grid with a good separation of the main topics: like Egypt's president Gamal Abdel Nasser and other articles about Arab countries, the Vietnam war, Charles de Gaulle, Germany, Nikita Khrushchev, a

war in Africa and so on. These units are expanded to provide a more detailed cluster representation at subsequent layers in the hierarchy. If we take a closer look at the 3x4 map below the unit labeled *viet* and *diem* (cf. Figure 4.4.2), we find articles about the Vietnam War in the upper half, and articles which contain information about Buddhism and the internal religious conflict in Vietnam in the lower half.

nasser	viet, diem
gaull, olympio	france, gaull
park	german
tshomb, park, death, track, combodia, ward, thailand, sarit, sihanouk, nkrumah	german, moscow, soviet, penkovsk
katanga, tshomb	moscow, khrushch

Figure 4.4.1: 1st Layer

viet	viet	viet
viet	viet, diem	
buddhist	buddhist, minh, diem	
buddhist		

Figure 4.4.2: 2nd Layer

4.5. Generative Topographic Mapping

Figure 4.5.1 shows the GTM mapping for the TIME magazine documents. Due to lacking a user interface support for a dataset with many (overlapping) clusters, we used the SOM results to analyze the GTM results. The clusters are basically identically with those found by the SOM.

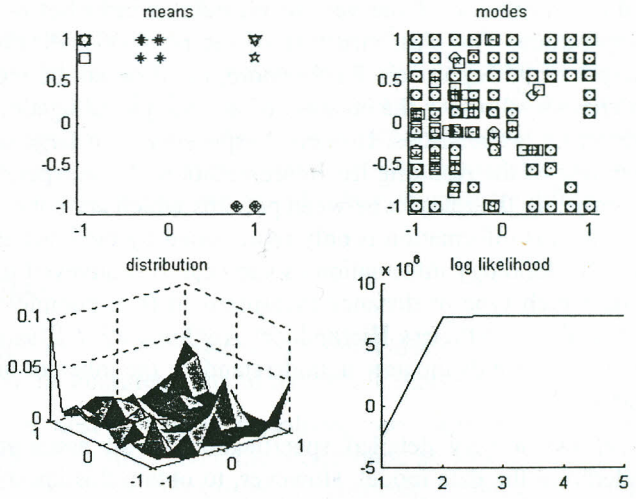

Figure 4.5.1: GTM TIME Magazine

Figure 4.5.1 depicts four diagrams. The first on the top left represents the means of the 420 single distributions. A few of them are marked with special symbols to identify them. Next to it the modes of the distributions are plotted. The modes and means can give some information on the types of the distributions. If the modes and means of one distribution are separated it indicates a multi-modal distribution. The diagram labeled distribution represents the density function of the complete data set. Peaks are

areas with a high density indicating clusters. Finally the log-likelihood is plotted. After only five cycles the batch algorithm converged.

The highlighted clusters in the means diagram are test clusters taken from the SOM results. For example, the cluster on the top right contains articles about Russia and NATO. As can be seen from the distribution diagram there is a peak there. Another peak can be found at the lower left, which is a cluster containing documents on the Vietnam War.

4.6. Comparison

Hierarchical Agglomerative Clustering with complete linkage, SOM, GHSOM and GTM find very similar clusters. AutoClass, even though not well adjusted to the dataset, also finds some of these clusters. One of the major differences was the computation time and storage used. AutoClass was most demanding, followed by GTM.

5. EVALUATION

The clustering qualities in principle were found to be similar, which can be attributed to the fact that all methods use similar assumptions about the underlying data distribution. However, we find that result representation and the information to be gained from these representations differs to a large degree.

Hierarchical Agglomerative Clustering can be used in a straightforward manner and provides a representation of the various clusters as branches of a tree, allowing very easy interpretation. The clear structure of the result visualization is one of the biggest advantages of this approach. Furthermore, the data can be viewed at different levels in the hierarchy, allowing the creation of and simple navigation through cluster structures of differing granularities. However, especially with large datasets it is easy to lose the overview of the resulting tree representation. An inexperienced user might have troubles analyzing the relation between patterns which are not close to each other in the hierarchy, as this information is only represented by the level at which the clusters are merged. No topology information as such can be conveyed using this method and the decision which type of distance measurement (e.g. complete linkage) to use might be non-trivial. Nevertheless Hierarchical Agglomerative Clustering is a simple and intuitive tool, which aids the user in understanding the inherent hierarchical structure in the dataset.

AutoClass allows a very detailed specification of all assumptions or a-priori knowledge underlying the data model. However, to obtain this knowledge, additional pre-processing steps such as analyzing the distribution across the individual attributes, have to be performed. Because of this wealth of possible parameter settings AutoClass demands a high understanding of statistics from the user. The results produced by AutoClass provide very detailed information on the probabilities of class assignments and class descriptions. However, no straightforward visualization of the results is available, requiring rather cumbersome manual interaction to elicit information from the result listings. AutoClass is a very powerful tool demanding a good understanding

of the clustering problem. However it is limited concerning the distributions of the attributes and the dimensionality of the data.

The remaining three methods, SOM, GHSOM and GTM are very similar as far as their result visualization possibilities are concerned. The main benefit of these methods is that they provide, some kinds of topological information in addition to the cluster information, i.e. they reveal additional information about the inherent structure of the data. This information would be difficult to abstract from the other methods. On the other hand it is necessary to be aware that abstracting the topology has a negative influence on finding the cluster structure itself. If no information about the topology is desired than other methods should be considered. The main usage of the topology is for visualization.

One of the disadvantages of the basic SOM method is the predefined, fixed size determining the granularity of the data representation in advance. Furthermore, no hierarchical structures can be detected from the basic SOM architecture. These limitations are being solved by the GHSOM, which automatically determines the size and a hierarchical structure within the dataset.

GTM, while being more precise, requires more computation time than the SOM. Especially with very big datasets the SOM offers big advantages over the GTM, since it is possible to use many algorithmic shortcuts, such as, for example, fast winner selection [8]. The higher precision of the GTM is mainly reflected in the way data items are assigned with a certain probability to a cluster or an area on the map, which is similar to AutoClass. Visualizing these probabilistic assignments is rather difficult, and with datasets with more than only a couple of items it is inevitable to resort to a SOM-like representation, partly losing this additional information again. This can be done as shown in Figure 4.3.1 by using the means of the single density functions. Since GTM, in contrast to the SOM, is based on a statistical framework, it can be proven to converge, has a theoretical basis for its parameters, and the a-posteriori probabilities can be used, as they are in AutoClass, to compare different models and to find the best.

6. CONCLUSION

We have empirically evaluated five different clustering algorithms with respect to their usability and the interpretability of their result representation.

SOM offers the most advantages visualizing very large and high dimensional datasets. If additional information about the hierarchical structure is desired, these can be obtained using the GHSOM. GTM offers a statistically sound visualization. AutoClass does not offer visualization, but it offers very accurate result descriptions. Hierarchical Agglomerative Clustering is a simple and very intuitive tool, aiding the user in understanding the hierarchical structure of the dataset. Especially in combination with other methods which do not reveal the hierarchical structure, it is very useful. In general, a more wide-spread combination and integrated result visualization of different approaches within one framework is highly recommended for complex data analysis tasks rather than parallel but separate representation.

REFERENCES

[1] M. C. Bishop, Svensén M., Williams C. K. I. (1996). GTM: a principled alternative to the self-organizing map. *Artificial Neural Networks – ICANN96 Proceedings of the 1996 International Conference, Lecture Notes in Computer Science*, vol. 1112, Springer Verlag.

[2] P. Cheesman, Stutz J. (1996). Bayesian Classification (AutoClass): Theory and Results. *Advances in Knowlege Discovery and Data Mining*, AAAI Press/MIT Press.

[3] P. Cheesman, Kelly J., Self M., Stutz J., Taylor W. and Freeman D. (1988). AutoClass: A Bayesian Classification System. *Proceedings of the 5th International Conference on Machine Learning*, Morgan Kaufman Publishers, San Francisco, 54-64.

[4] G. Deboeck and Kohonen T. (1998). *Visual Explorations in Finance*. Springer Verlag, Berlin, Germany, 1998.

[5] M. Dittenbach, Merkl D. and Rauber A. (2000). The Growing Hierarchical Self-Organizing Map. *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2000)*.

[6] B. Fritzke (1996). Growing Self-Organizing Networks – Why? *Proceedings ot the European Symposium on Artificial Neural Networks (ESANN96)* Bruges, Belgium.

[7] A. K. Jain, Murty M. N. and Flynn P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys*, Vol 31, No. 3, 264-323.

[8] S. Kaski (1999). Fast winner search for SOM based monitoring and retrieval of high dimensional data. *Proceedings of ICANN99, 9th International Conference on Artificial Neural Networks*, IEE, vol. 2, London, 940-945.

[9] B. King (1967). Step-wise clustering procedures. *Journal of the American Stat. Assoc.* 69, 86-101.

[10] T. Kohonen (1982). *Self-Organized formation of topologically correct feature maps*. Springer.

[11] T. Kohonen (1997). *Self-Organizing Maps*.Second Extended Edition, Springer Verlag.

[12] D. Merkl and Rauber A. (1997). Finding Structure in Text Archives. *Proceedings of the European Symposium of Artificial Neural Networks (ESANN98)*.

[13] D. Merkl and Rauber A. (1998). Cluster Connections – A visualization technique to reveal cluster boundaries in self-organizing maps. *Proceedings of the 9th Italian Workshop on Neural Nets (WIRN97)*. Springer.

[14] R. Miikkulainen (1990). *Script Recognition with Hierarchical Feature Maps*. Connection Science 2.

[15] Rauber, A. (1999. LabelSOM: On the Labeling of Self-Organizing Maps. *Proceedings of the International Joint Conference on Neural Networks (IJCNN99)*.

[16] A. Rauber, Dittenbach M., Merkl D. (2000). Automatically Detecting and Organizing Documents into Topic Hierarchies: A Neural-Network Based Approach to Bookshelf Creation and Arrangement. *European Conference on Research and Development for Digital Libraries (ECDL'00)*, Lisboa, Portugal, 2000, Lecture Notes, Springer Verlag.

[17] H. Ritter and Kohonen T. (1989). Self-Organizing Semantic Maps. *Biological Cybernetics* 61, 241-254. Springer.

[18] O. Simula, Vasara P., Vesanto J. and Helminen R. (1999). The Self-Organizing Map in Industry Analysis. In: Jain, L. and Vemuri, V. (eds.): *Industrial Applications of Neural Networks*, CRC Press, Washington, DC.

[19] P.H.A. Sneath and Sokal R. R. (1973). *Numerical Taxonomy*. Freeman London UK.

[20] A. Ultsch (1993). *Self-Organizing Neural Networks for Visualization and Classification*. In: Information and Classification. Concepts, Methods and Applications, Springer, 1993.

Received: 16 Mart 2000
Accepted: 19 May 2000

Andreas Rauber
Jan Paralic
Elias Pampalk

EMPIRIJSKA PROCJENA ALGORITAMA GRUPIRANJA

Sažetak

Nenadzirana klasifikacija podataka može se smatrati jednim od najvažnijih početnih koraka u postupku rudarenja podataka. Mnogi algoritmi razvijeni su i koriste se u ovom kontekstu u raznim područjima primjene, iako je dostupno samo malo dokaza za to koji algoritmi bi se trebali koristiti u kojem kontekstu i koje tehnike nude obećavajuće rezultate kad se kombiniraju za određeni zadatak. U ovom radu predstavljamo empirijsku procjenu nekih istaknutih nenadziranih tehnika klasifikacije podataka s obzirom na njihovu upotrebljivost i mogućnost interpretacije njihovog prikaza rezultata.

Ključne riječi: rudarenje podataka, analiza grupe, hijerarhijsko aglomeracijsko grupiranje, Bayesovo grupiranje, samoorganizirajuća karta (SOK), rastuća hijerarhijska SOK, generativno topografsko preslikavanje.