Cluster Analysis and Data Mining Applications

Thomas A. Runkler

Siemens AG Corporate Technology
Information and Communications
81730 Munich, Germany

Phone: ++49 89 636 45372
Fax: ++49 89 636 45456
Email: Thomas.Runkler@mchp.siemens.de

ABSTRACT: The data mining process involves various steps like data preparation, preprocessing, pattern recognition, and postprocessing. For pattern recognition statistical as well as soft computing methods such as fuzzy logic, neural networks, or machine learning are applied. The advantage of clustering methods for pattern recognition in the data mining context is that there are powerful families of clustering algorithms available for a wide variety of different prototype shapes or membership functions. Moreover, fuzzy clustering results can directly be interpreted as expert system rules leading to an easy evaluation of the data mining results. Four selected data mining applications from Siemens Corporate Technology show that these methods are good choices for real world applications.

I. INTRODUCTION: THE DATA MINING PROCESS

The information in the world doubles about every 20 months. Among the most rapidly growing sources of data are (besides the internet) industrial process control systems, commercial data bases, biotechnology, and automatic imaging systems. Data collection and storage becomes more and more difficult with growing sizes of the data bases. The biggest challenges, however, seem to be the automatic processing, evaluation, analysis, and interpretation of the data.

"Data mining" methods try to extract the "knowledge", i.e. the interesting patterns contained in the data sets [6]. Patterns are considered interesting if they are general, non–trivial, new, useful, and understandable. To find patterns in data conventional statistical methods like correlation and regression analysis are applied as well as soft computing methods like clustering, fuzzy logic, neural networks, or machine learning [5].

In practical applications the data sets are often not immediately available, and the pattern recognition results often do not immediately solve the problems that data mining is used for. Therefore, additional data preparation, preprocessing, and postprocessing steps are necessary. The main steps of data preparation are planning and performing the data collection, generating meaningful features, and selecting the relevant parts of the data. The main steps of preprocessing are normalizing, cleaning, and filtering the data, completing missing data, correcting erroneous data, and transforming the data to the appropriate coordinate systems. The main steps of postprocessing are the interpretation of the pattern recognition results, the documentation of the newly discovered knowledge, and the evaluation and validation of the new information.

The overall structure of this data mining process is visualized in Figure 1. Clearly, not all of these steps necessarily have to be performed in every data mining project. The methods mentioned here are just typical representatives of a data mining tool box, and for each application it is necessary to pick to most appropriate tools out of it.

II. DATA MINING WITH CLUSTERING

Since many of the data mining methods shown in Figure 1 are well–known, I would like to focus here on cluster analysis methods [8]. Cluster analysis has proven to be a powerful tool to detect structure in data.

Cluster structure can be found in data by sequential, objective function, or cluster estimation methods. Among the *sequential* methods we have sequential agglomerative hierarchical non–overlapping (SAHN) and sequential divisive hierarchical non–overlapping (SDHN) clustering [13].

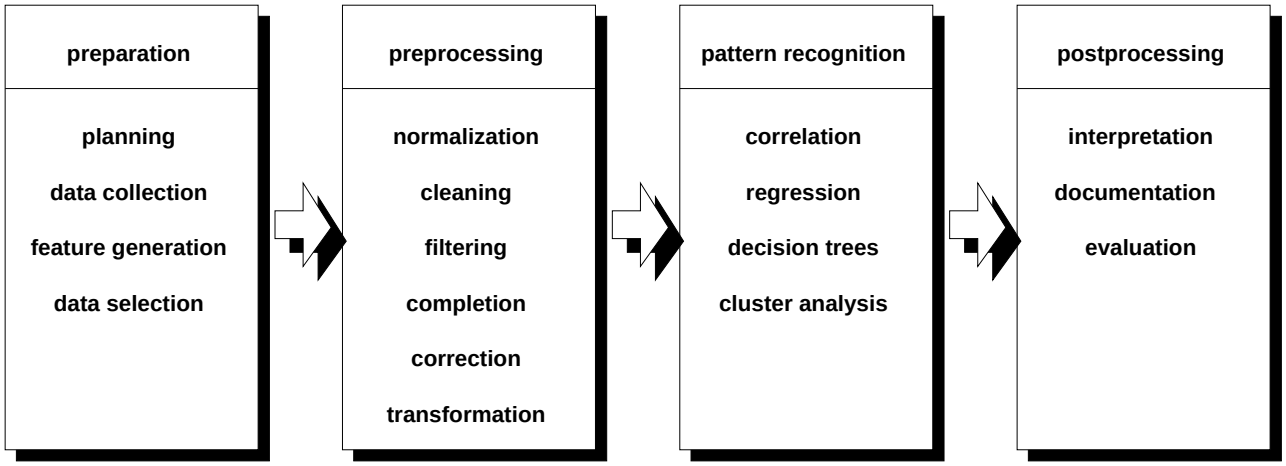

Figure 1. The most important processing steps in the data mining process.

The most widely known *objective function* methods for clustering are extensions of c–means models. Ball and Hall [1] developed a crisp c–means model that detects c cluster centers in data and assigns each data vector to one of the c clusters. Bezdek [2] extended this model to the fuzzy case. The fuzzy c–means model uses a fuzzy partition, i.e. each data vector can belong to several of the c clusters to a certain extent, which is quantified by a number beween 0 and 1; the sum of the memberships of each data vector to all the clusters is equal to one; and there is no empty cluster. These clustering models for point prototypes can be extended to higher order prototypes like in the Gustafson–Kessel [7], the c–lines [3], or the c–elliptotypes model [4]. Virtually, any prototype shape like circles or shells can be detected using this family of clustering models. This is an important property for data mining!

All models of the c–means family determine local extrema of objective functions. This is mostly done by alternating optimization using the first order necessary conditions for the extrema (Picard iteration). This alternating optimization leads to an algorithmic structure that initializes cluster prototypes first and then alternatingly computes memberships and prototypes. The equations to do that are directly derived from the objective functions. They can be interpreted then as membership and prototype functions. Runkler and Bezdek [12] therefore abandoned the objective functions but specified the clustering algorithms directly by their (application specific) membership and prototype function. The resulting *alternating cluster estimation (ACE)* model thus allows the use of virtually any membership and prototype function.

A big advantage of clustering for data mining is that the clusters can be interpreted as rules. Depending on the type of prototypes used fuzzy clusters can be directly translated into Mamdani–Assilian rules [9] like

$$R_{im}: \text{ IF } \bigwedge_{l=1}^{p} \mu_{il}(x^{(l)}) \text{ THEN } \nu_{im}(y^{(m)}) = \mu_{i\,(p+m)}(y^{(m)}),$$

into Sugeno–Yasukawa [16] rules like

$$R_i: \text{ IF } \bigwedge_{l=1}^{p} \mu_{il}(x^{(l)}) \text{ THEN } y = y_i,$$

or into Takagi–Sugeno rules [17] like

$$R_i: \text{ IF } \bigwedge_{l=1}^{p} \mu_{il}(x^{(l)}) \text{ THEN } y = f_i(x).$$

Extracted knowledge in the form of these rules can be quite easily interpreted by humans.

III. DATA MINING APPLICATIONS

Data mining is being applied at Siemens Corporate Technology in various industrial projects. Here, I want to mention just four typical examples.

Clustering was used to analyze the production process in a paper factory [10]. The quality of the paper produced there strongly depends on a large number of different chemical and physical variables, such as concentrations of

different chemicals, pressures, temperatures, densities, specific weights, and many more. The 27 most important variables were selected after discussions with the plant experts. A data set with 2594 vectors of these variables was used to detect the most important influence factors on the paper quality. As a result, one specific weight was identified as the critical variable. The influence of this variable on the paper quality is clearly nonlinear, so conventional linear statistics would not have been able to detect this dependency. After this investigation the production process could be optimized according to the newly gained knowledge. After this optimization the paper mill works more efficiently. It saves money, energy, and material.

In a second application, mammographic images had to be preprocessed for the detection of breast cancer [11]. One of the main tasks in medical image preprocessing is the detection of the region of interest. This is an image segementation task. Each pixel has to be assigned to a certain class, for example to the tissue, bone, skin, or background. A close analysis of the generated features discovered fractal–like linear structures. The lines in this fractal–like feature space were discovered by fuzzy c–elliptotypes clustering. The resulting classifier was successfully applied to image segmentation and enhancement.

The third application tries to detect traffic jams from data obtained by inductive coils on German autobahns [14]. The main problem here is that the physical relations between the traffic parameters flux and density depend considerably of external conditions like the weather which can not be measured appropriately. Our data mining solution to this problem detects various specific patterns like line segments and elliptic regions in the inductive coil data using fuzzy c–mixed prototypes clustering. The positions, shapes, and directions of these patterns reflect the current traffic situation and allow an accurate traffic classification even under strongly varying external conditions.

The final application presented here is the long–term traffic prediction [15] based on data obtained by inductive coils in German cities plus information about specific external conditions like the day of the week, holidays, or fair seasons. Fuzzy cluster analysis of these data yielded a knowledge base that is implemented as a fuzzy expert system. This expert system was successfully used to predict urban traffic characteristics.

IV. CONCLUSIONS

The application examples for data mining using different clustering methods show that these methods are well–suited for real world applications. The main advantages of the clustering methods are that they can be easily used with a large scope of different prototypes and membership functions, and that the clustering results can be easily written in the form of if–then rules. In the data mining context this means that a wide variety of different patterns can be automatically detected from large data sets, and that even complicated data mining results can be easily interpreted an evaluated.

REFERENCES

[1] G. H. Ball and D. J. Hall. Isodata, an iterative method of multivariate analysis and pattern classification. In *Proc. IFIPS Congress*, 1965.

[2] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York, 1981.

[3] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson. Detection and characterization of cluster substructure, I. Linear structure: Fuzzy c–lines. *SIAM Journal on Applied Mathematics*, 40(2):339–357, April 1981.

[4] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson. Detection and characterization of cluster substructure, II. Fuzzy c–varieties and convex combinations thereof. *SIAM Journal on Applied Mathematics*, 40(2):358–372, April 1981.

[5] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1974.

[6] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining.* AAAI Press, Menlo Park, 1996.

[7] E. E. Gustafson and W. C. Kessel. Fuzzy clustering with a covariance matrix. In *Proc. IEEE CDC, San Diego*, pages 761–766, 1979.

[8] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis — Methods for Image Recognition, Classification, and Data Analysis.* Wiley, 1999.

[9] E. H. Mamdani and S. Assilian. An experiment in linguistic synthesis with a fuzzy logic controller. *International Journal of Man–Machine Studies*, 7(1):1–13, 1975.

[10] T. A. Runkler. Automatic generation of first order Takagi–Sugeno systems using fuzzy c–elliptotypes clustering. *Journal of Intelligent and Fuzzy Systems*, 6(4):435–445, 1998.

[11] T. A. Runkler and J. C. Bezdek. Image segmentation using fuzzy clustering with fractal features. In *Proc. IEEE International Conference on Fuzzy Systems*, volume 3, pages 1393–1398, Barcelona, July 1997.

[12] T. A. Runkler and J. C. Bezdek. Alternating cluster estimation: A new tool for clustering and function approximation. *IEEE Transactions on Fuzzy Systems*, to appear, 1999.

[13] P. Sneath and R. Sokal. *Numerical Taxonomy*. Freeman, San Francisco, 1973.

[14] C. Stutz and T. A. Runkler. Fuzzy c–mixed prototype clustering. In W. Brauer, editor, *Fuzzy–Neuro–Systems '98, München*, volume 7 of *Proceedings in Artificial Intelligence*, pages 122–129, March 1998.

[15] C. Stutz and T. A. Runkler. Classification and prediction of road traffic using application-specific fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, submitted, 1999.

[16] M. Sugeno and T. Yasukawa. A fuzzy–logic–based approach to qualitative modeling. *IEEE Transactions on Fuzzy Systems*, 1(1):7–31, February 1993.

[17] T. Takagi and M. Sugeno. Fuzzy identification of systems and its application to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, 15(1):116–132, 1985.