LA - A Clustering Algorithm with an Automated Selection of Attributes, which is Invariant to Functional Transformations of Coordinates

Mikhail V. Kiselev[1], Sergei M. Ananyan[2] , and Sergey B. Arseniev[1]
[1] Megaputer Intelligence Ltd., 38 B.Tatarskaya, Moscow 113184 Russia
{M.Kiselev, S.Arseniev}@megaputer.com
http://www.megaputer.com
[2] IUCF, Indiana University, 2401 Sampson Lane, Bloomington, IN 47405 USA
sananyan@indiana.edu

Abstract. A clustering algorithm called LA is described. The algorithm is based on comparison of the n-dimensional density of the data points in various regions of the space of attributes $p(x_1,...,x_n)$ with an expected homogeneous density obtained as a simple product of the corresponding one-dimensional densities $p_i(x_i)$. The regions with a high value of the ratio $\frac{p(x_1,...,x_n)}{p_1(x_1)...p_n(x_n)}$ are considered to contain clusters. A set of attributes which provides the most contrast clustering is selected automatically. The results obtained with the help of the LA algorithm are invariant to any clustering space coordinate reparametrizations, i. e. to one-dimensional monotonous functional transformations $x' = f(x)$. Another valuable property of the algorithm is the weak dependence of the computational time on the number of data points.

1 Introduction

Clustering is one of the typical problems solved by data mining methods [5]. This is the process of grouping cases or database records into subsets such that the degree of similarity between cases in one group is significantly higher than between members of different groups. An exact definition of the similarity between cases, as well as other details, vary in different clustering methods. Most often used algorithms can be roughly associated in the following groups.

1. Joining methods. In these methods smaller clusters are consequently merged in larger clusters.

2. K-means methods [2]. These methods find an *a priori* specified number of clusters such that variation of attribute values inside clusters would be significantly less than variation between clusters. In order to increase the clustering significance (to decrease the respective p-value) data points are exchanged between clusters.

3. Seeding algorithms [10]. In these methods a certain number of initially selected data points serve as the seeds for growing clusters.

4. Density-based algorithms. The space of attribute values is broken into a set of regions. The regions which have significantly higher point density are considered as containing clusters of data points.

J.M. Zytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 366-371, 1999.

5. Algorithms based on neural networks [1, 4, 8, 11].

Yet, despite the variety of the approaches and methods, practical data mining problems require a further improvement of clustering algorithms. In our opinion, many modern clustering algorithms have the following weak sides:

1. **High computational complexity**. The computational time of many clustering algorithms depends on the number of records at least as $O(N^2)$ (see parallel realization of clustering algorithms in [9]).

2. **Insufficient performance with multi-dimensional data**. In databases where every record contains a large number of numerical, boolean and categorical fields the right choice of attributes for the clustering procedure often determines the quality of the result obtained. An automated selection of several attributes most crucial for clustering out of, say, hundreds of fields present in the database would be a very desirable feature for clustering algorithms implemented in a data mining system. Yet only a few of existing algorithms offer such a possibility.

3. **Sensitivity to functional transformations of attributes**. Suppose we would like to find clusters in a database describing the customers of some retailer. Every customer is described by her or his age and monthly income. These variables are measured in different units. Since many clustering algorithms use euclidean metrics which in our case can be written as $dist(R_1, R_2) = \sqrt{A(age_1 - age_2)^2 + (income_1 - income_2)^2}$, different choice of the constant A would give us a different set of clusters. Besides, it is evident that clustering performed in terms of *(age, log(income))* instead of *(age, income)* leads in general to completely different results.

4. **Lack of effective significance control**. The clustering procedures implemented in many existing data mining systems and statistical packages find clusters even in the data consisting of artificially generated random numbers with a uniform distribution. It would be highly desirable that clusters found by these systems express objective and statistically significant properties of data - not simply the statistical fluctuations [3].

In the present paper we describe a clustering algorithm called LA (the abbreviation stands for Localization of Anomalies - point density anomalies are implied), which is free of the drawbacks listed above.

2 Automated Clustering of Database Records Including Multiple Fields

Prior to discussing our algorithm we say a few words about our understanding of the term "cluster". In many approaches a set of clusters found by the corresponding algorithm should be considered as a property of the concrete dataset which was explored. An individual cluster is characterized completely by the set of datapoints that belong to it. We consider a cluster as a region in the space of attribute values which has a significantly higher concentration of datapoints than other regions. Thus, it is described mainly by boundaries of this region and it is assumed that other sufficiently representative datasets from the universum of data belonging to the same

application domain will also have a higher density of points in this region. Therefore the discovered set of clusters may not include all the records in the database. Beside that, the problem of the determination of the statistical significance of clustering becomes very important.

In our approach each cluster is represented as a union of multi-dimensional rectangular regions described by a set of inequalities $x < a$ or $x \geq a$ for numerical fields x and by a set of equalities $c = A$ for categorical fields c.

Our algorithm is applied to a database DB which can be logically represented as a rectangular table with N rows and M columns. This set of attributes (columns) will be denoted as **A**. We consider databases with numerical fields only. The extension of this method to categorical variables is quite evident. Thus, database DB can be represented as a finite set of points in the M-dimensional space $\Re^M$. Coordinates in $\Re^M$ will be denoted as x_i, $i=1,...,M$.

The LA algorithm consists of two logical components. The purpose of the first component is the selection of the best combination of attributes x_i which gives the most significant and contrast clustering. The second component finds clusters in space of a fixed set of attributes x_i. We begin our consideration with the second part.

Suppose that we fix m attributes from M attributes presented in the database DB. Our approach is based on breaking the space of attribute values $\Re^m$ in a certain set of regions $\{E_i\}$ and comparing the density of points in each region E_i. Namely, we cut $\Re^m$ by hyperplanes x_i = const and take the rectangular regions formed by these hyperplanes as E_j. We call such set of regions the grid $\{E_i\}$. The hyperplanes forming the grid may be chosen by various methods. However it is important that datapoints would be distributed among the cells E_i as evenly as possible.

Consider one cell E_i. Let n be the number of datapoints in this cell. The cell E_i can be considered as a direct product of the attribute axes segments: $E_i = S_1 \times \ldots \times S_m$. Let us denote the number of points with the value of the j-th attribute falling into the segment S_j as M_j. If the points do not form clusters in the space of attributes x_i which are considered as independent then the relative density of points in E_i, is approximately equal to multiplication of one-dimensional relative densities of points in segments S_j:

$$\frac{n}{N} \approx p_j = \frac{M_1 \ldots M_m}{N^m} \tag{1}$$

A significantly higher value of $\frac{n}{N}$ would mean that E_i should be considered as (a part of) a cluster. In the case of $m = 1$ the approximate equality (1) is trivially exact. Thus the minimum dimension of the clustering space m is 2. To find clusters consisting of rectangular regions with anomalous point density we use the following procedure.

For each cell E_i with the number of points greater than $Np_i = \frac{M_1 \ldots M_m}{N^{m-1}}$ we calculate the probability that the high density of points in this cell is a result of the statistical fluctuation. Namely, we determine for each cell E_i the value of

$s_i = b(n, N, \frac{M_1 \ldots M_m}{N^m}) = b(n, N, p_i)$ where *b(k, K, p)* is a tail area probability of the binomial distribution with the number of trials *K* and the event probability *p*. A list of all E_i ordered by ascending values of s_i is created. Denote the ordered sequence of the cells as $\{ E'_j \}$. For each cell E'_j we know the number of points lying in the cell, n_j, and the value of p_j. For each *j* we calculate value $s_{CUM_j} = b(\sum_{i=1}^{j} n_i, N, \sum_{i=1}^{j} p_i)$. Let us denote the value of *j* for which s_{CUM_j} is minimal as j_{BEST}; this minimum value of s_{CUM_j} will be denoted as s_{BEST}. This value corresponds to the most contrast, most significant division of all cells E_i into "dense" and "sparse" ones. Let us consider the cells E'_j with $j \le j_{BEST}$. In this set of cells we search for subsets of cells C_k such that all of them satisfy the following conditions: 1) either the subset C_k contains only one cell or for each cell *E* belonging to the subset C_k there exists another cell in C_k which has a common vertex or border with cell *E*; 2) if two cells belong to different subsets they have no common vertexes or borders. We call these subsets clusters.

Thus, for each subset **a** of attributes $\mathbf{a} \subset \mathbf{A}$, $|\mathbf{a}| = m$ satisfying the condition (1) we can determine a set of clusters **C**(**a**), the clustering significance $s_{BEST}(\mathbf{a})$, and the total number of points in all clusters $K(\mathbf{a})$. Now let us discuss the procedure which selects the best combination of attributes for clustering. The purpose of this procedure is finding a subset of attributes which has the maximum value of some criterion. In most cases it is natural to choose 1 - s_{BEST} as such a criterion. Other possible variants are the number of points in clusters or the number of clusters. It is often required that the clustering procedure should elicit at least two clusters and also that 1 - s_{BEST} should be greater than a certain threshold confidence level. It is obvious that in order to satisfy the first requirement each coordinate should be divided in at least three sections. Depending on the actual conditions of the data exploration carried out (possible time limitation) various modifications of the procedure can be utilized. We consider two extreme cases.

a. Full search. All combinations of *m* attributes ($1 < m \le \frac{1}{2}\log_3 N$) are tried. The best combination is selected.

b. Linear incremental search.

Step 1. All combinations of two attributes are tried. The best pair is included in list of selected attributes **SEL**. The respective value of the criterion will be denoted as $R(\mathbf{SEL})$.

Step 2. If $|\mathbf{SEL}| > \frac{1}{2}\log_3 N$ or **SEL** includes all attributes the process stops and **SEL** is the result.

Step 3. All combinations of attributes consisting of all the attributes from **SEL** plus one attribute not included in **SEL** are tried. Let the best combination be

$\mathbf{SEL}' = \mathbf{SEL} \cup \{a\}$. If $R(\mathbf{SEL}') \le R(\mathbf{SEL})$ the process stops and $\mathbf{SEL}$ is selected as a final set of attributes.

Step 4. Set $\mathbf{SEL} = \mathbf{SEL}'$ and go to Step 2.

An abundance of intermediate variants of this procedure can be constructed.

3 Properties of LA Algorithm

It can be easily proven that the considered LA algorithm has the following properties:

1. If we replace a numerical attribute x with its functional derivative $f(x)$, where f is a monotonous function and use $f(x)$ instead of x, this will not change the clustering results. The algorithm will detect the same number of clusters and the same sets of records will enter the same clusters.

2. The computational time depends on the number of records N only weakly. The measurements show that the most time consuming operation is the sorting of the values of attributes when the grid $\{E_i\}$ is constructed. This operation requires $O(mN\log N)$ time. The exact computational time of LA algorithm depends on the version of the procedure used for selecting the best attributes. One can see that for a fast linear search the computational time is $O(M^3 N\log N)$.

3. The LA algorithm works best in the case of a great number of records. The less records are explored, the less fine cluster structure is recognized. In the worst case, when a cluster of the size approximately equal to one cell is intersected by a hyperplane it may not be detected by the algorithm.

4. The LA algorithm is noise tolerant. Indeed, the algorithm is based not on the distances or other characteristics of single points but on the properties of substantial subsets of data. Thus an addition of a relatively small subpopulation of points with different statistical properties ("noise") cannot influence the results obtained by the algorithm substantially.

4 Conclusion

We have described a new algorithm for finding clusters in data called LA. At present the LA algorithm is implemented as a data exploration engine in the PolyAnalyst data mining system [6, 7]. Our algorithm can select automatically an optimal subset of the database fields for clustering. The algorithm is invariant to a monotonous functional transformation of numerical attributes and has a weak dependence of the computational time on the number of records in the database. The algorithm is based on the comparison of the n-dimensional density of the data points in various regions of the space of attributes with an expected homogeneous density obtained as a simple product of the corresponding one-dimensional densities. As a part of PolyAnalyst system it has been practically used in the fields of database marketing and sociological studies.

References

1. Carpenter, G. and Grossberg, S. A Massively Parallel Architecture for a Self-Organizing Neural Pattern Recognition Machine, Computer Vision, Graphics, and Image Processing, 37:54-115, 1987.
2. Cheng, Y. Mean shift, mode seeking, and clustering, IEEE Trans. on Pattern Analysis and Machine Intelligence, 17:790-799, 1995.
3. Dave, R.N., Krishnapuram, R. Robust clustering methods: a unified view, IEEE Trans. on Fuzzy Systems, 5:270-293, 1997.
4. Hecht-Nielsen, R. Neurocomputing, Reading, MA: Addison-Wesley, 1990.
5. Jain, A. K. and Dubes, R. C. Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.
6. Kiselev, M.V. PolyAnalyst 2.0: Combination of Statistical Data Preprocessing and Symbolic KDD Technique, In: Proceedings of ECML-95 Workshop on Statistics, Machine Learning and Knowledge Discovery in Database*s*, Heraklion, Greece, pp. 187-192, 1995.
7. Kiselev, M.V., Ananyan, S. M., and Arseniev, S. B. Regression-Based Classification Methods and Their Comparison with Decision Tree Algorithms, In: Proceedings of 1st European Symposium on Principles of Data Mining and Knowledge Discover*y*, Trondheim, Norway, Springer, pp 134-144, 1997.
8. Kohonen, T. Self-Organizing Maps, Berlin: Springer-Verlag, 1995.
9. McKinley, P. K. and Jain A., K. Large-Scale Parallel Data Clustering, IEEE Trans. on Pattern Analysis and Machine Intelligence, 20:871-876, 1998.
10. Milligan, G.W. An estimation of the effect of six types of error perturbation on fifteen clustering algorithms, Psychometrika*,* vol 45, pp 325-342, 1980.
11. Williamson, J. R. Gaussian ARTMAP: A Neural Network for Fast Incremental Learning of Noisy Multidimensional Maps. Technical Report CAS/CNS-95-003, Boston University, Center of Adaptive Systems and Department of Cognitive and Neural Systems, 1995.