УДК 519

B. Mirkin, M. Levin, E. Bakaleinik
School of Computer Science and Information Systems
Birkbeck College

Intelligent K-Means Clustering in Analysis of Newspaper Articles on Bribing

В работе предлагается дополнить метод к-средних – один из самых популярных методов кластер-анализа – средствами: а) предварительной обработки данных; б) задания начальных центроидов и их количества; в) интерпретации полученных результатов. Эти дополнения вытекают из результатов Миркина (1999, 2001). В частности, специальные методы разработаны для поиска начальных центроидов (Разделяй и Властвуй) и отыскания отличительных логических описаний кластеров (Аппкод). Полученная таким образом модификация метода к-средних, называемая интеллигентный к-средних, применяется к анализу 55 статей, опубликованных в центральной российской прессе об отдельных случаях взяточничества. Полученные результаты указывают, что все многообразие публикаций может быть резюмировано так: в каждой из основных отраслей – разные виды коррупции: в региональной администрации – вымогательство и протекция, в органах правопорядка – нарушение законности и прикрывательство, в других – изменение категории. Серьёзная проблема, ждущая своего разрешения – разработка методов формирования значимых признаков для описания того или иного корпуса текстов.

Introduction

K-Means is a most popular clustering algorithm because of its undisputed advantages over other techniques. Still, there are few aspects of the algorithm, which so far have received relatively little attention in the literature. These aspects include issues of initial setting and interpretation aids, leaving them to imagination and fantasy of the user. In this paper, we present model-based recommendations on both of these items following Mirkin (1999, 2001). We describe Separate/Conquer version of K-Means as a tool for finding an initial setting for K-Means. The interpretation aids are derived from the data scatter decomposition due to a cluster structure found. They can be presented at three levels: cluster representative, cluster tendency, and cluster description.

The method is illustrated with a relatively small real-world data set extracted from articles on corruption cases published in Russian newspapers. The analysis leads us to conclude that different branches of administration (government, law enforcement, other) involve different types of corruption services. The results show that the intelligent K-Means can be used for semantic annotation of media data. However, to make it effective, the problem of automatically extracting meaningful features from media items should be addressed.

1. What Is Intelligent K-Means

1.1. Straight K-Means and Its Properties

K-Means is a major clustering technique that is present, in various versions, in statistical packages such as SPSS, SAS and SYSTAT and data mining programs such as Clementine and MineSet. Information on these packages can be found on the web.

The algorithm processes a data set presented in the entity-to-feature format and produces a set of non-overlapping clusters of entities along with cluster centroids that are "model" entities with within-cluster averaged features. The algorithm reiterates the same two-step computation until the result is stabilised. At the first step, given cluster centroids, the algorithm updates assigning of each of the entities to that of centroids which is closest to the entity. At the second step, given cluster lists, the within cluster means of all features are calculated and put as updated cluster centroids.

There are many advantages in this method. In particular:

1. It models typology building (via centroids);
2. It is computationally effective both in memory and time;
3. It can be utilised incrementally, «on-line».

Somewhat less known are properties highlighted in Mirkin (1999, 2001):

4. It straightforwardly associates feature salience weights with feature scales;
5. It can be applied to both quantitative and categorical data as well as mixed data, provided that care has been taken of the relative feature scaling.

Less attractive properties of the generic K-Means:

6. Simple convex spherical shape of clusters in the feature space;
7. Instability of the results with respect to initial setting (the number of clusters and initial centroids).
8. Insufficient built-in interpretation aids.

Item 6 implies that the feature set should be chosen carefully so that centroids and spheres around them could be indeed interpreted as certain types concurring with goals of the data analysis. Item 7 implies that the initial setting, in fact, much affects the solution and thus must be carefully selected based on conceptual understanding of the knowledge domain or preliminary data analyses. Item 8 implies the need in theoretical analysis of the assumptions underlying the method and deriving computational methods advancing into interpretation issues.

What are the options facing the laymen user who has an embryonic knowledge of the domain? More studies and experiments? This is in most cases not quite a practical advice. Sometimes a more viable strategy would be of a better usage of properties of the method. Intelligent K-Means is a set of tools providing the user with instruments for automatically tackling the issues of initial setting and interpretation aids.

1.2. Separate/Conquer Version of K-Means to Set Initial Centroids

The Separate/Conquer method (Mirkin, 1999) utilises the concept of reference point that is what the user considers as the normal pattern of features, typically, the grand mean of the entity set. Having the reference point specified, Separate/Conquer finds a cluster that represents the pattern, which is most deviant from the reference

point. This may be used for finding anomalous patterns, but here is considered only as a tool for preliminary analysis.

In the beginning, Separate/Conquer puts the initial centroid of the deviant cluster at the entity which is most distant from the reference point and then reiterates K-Means steps with regard to the centroid and the reference point. Given a centroid, all entities are assigned to either the centroid or reference point depending on the distances to them. Then, the centroid is updated as the average of all entities assigned to it. It is guaranteed that after a number of steps, the centroid doesn't change anymore, which completes the process of finding the deviant cluster. With this cluster removed from the data set, the next cluster is found with the same algorithm. This goes on until no non-clustered entities remain. There can be other stopping criteria as well. The experiments and the underlying theory show that the reference point should remain unchanged during the entire process of extracting of clusters.

This method works especially well when the data may be thought of as a collection of clusters of various sizes located at different distances from the reference point. Some clusters extracted with Separate/Conquer method may be singletons or doubletons exposing some strange patterns that can be associated with outliers generated by different causes such as errors. Having such small clusters removed, the other clusters can be considered as a fair representation of the data structure; the number of clusters and their representatives are put as the initial setting for K-Means in the suggested version of intelligent clustering.

This `intelligence' is based on a specific type of cluster structure assumed. In our opinion, no intelligence is possible without an assumption of underlying structure of the world at hand.

1.3. Interpretation Aids

Typological description of a cluster can be done on either or all of the following levels:

1. By pointing to its typical representative or prototype. These two do not necessarily coincide as they may express different aspects. The former, a typical representative, illustrates average pattern of features in the cluster. The latter, a prototype, may rather relate to those features that separate the cluster from the rest and thus may get distorted towards to them. Both can be of interest to the user such as a planning council or marketing unit that is going to operate with the entities.

2. By describing the average tendencies of the cluster entities in terms of the most salient features. This can be of interest to the user who tries to generalise on the structure of the domain represented by the data set and capture those tendencies that describe and, potentially, explain the structure.

3. By distinctively describing the cluster in terms of the most salient features. This can be used by those who want to set a knowledge item with respect to the pattern captured in the cluster or/and develop a definitive decision rule to distinguish cluster entities and the like from the rest. These two aims do not necessarily coincide. The knowledge discovery should involve such features that can be supported within a theoretical framework, while decision rules may utilise some «handy» features that happen to work well just within a limited scope of the decision making process.

The model underlying K-Means clustering can be utilised for deriving interpretation aids at any of these levels. The basic equation has the format of the additive decomposition of the data scatter into explained and unexplained parts:

$$\sum_{i=1}^{N}\sum_{v=1}^{V} y_{iv}^2 = \sum_{k=1}^{K} c_{kv}^2 N_k^2 + \sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v=1}^{V} (y_{iv} - c_{kv})^2 \quad (*)$$

where N is the number of entities, V of variables and K the number of clusters that are denoted by i, v, and S_k, respectively. Denotations y_{iv} and c_{kv} refer to pre-processed values of variable k at entity i and centroid k, respectively, and N_t denotes the cardinality of cluster S_k. It is important, in the present context, that equation (*) assumes that the data pre-processing has been done by shifting the space origin into the reference point and normalising the variables by their theoretical or empirical ranges.

Equation (*) leads to the following recommendations with regard to interpretation aids:

1. The typical representative of cluster k is an entity that is the closest to centroid vector c_k (by Euclidean distance) and the prototype is an entity whose scalar product with c_k is maximal.

2. The salience of a feature v at cluster k is proportional to the squared difference of grand mean and within-cluster mean of v, expressed in the value of c^2_{kv}. The tendencies are expressed in centroid values c_{kv} of most salient features.

3. Distinctive descriptions of clusters as conjunctions of statements of the format "feature v's range is between av and bv" where v, av and bv are chosen according to the ordering of the salience weights with forward and backward sequential searches (see algorithm APPCOD (Mirkin, 1999, Mirkin & Ritter, 2000)).

These will be illustrated in the remainder.

2. Analysis of Newspaper Articles on Bribing

2.1. Description of the data

It appears, there are not many articles in Russian newspapers about bribing containing enough detail. We collected corruption related articles published in central newspapers from June 1999 through May 2000. Most of them have come from newspaper Kommersant Daily.

To derive a data set from our collection, we consider that there are five structural aspects in any corruption case: (a) official side, (b) client side, (c) service provided, (d) corrupt interaction, and (e) environment. These structural aspects can be characterised by the following eleven features that can be recovered from newspaper articles:

(a) Office

1. Level of office
   - enterprise
   - city
   - regional
   - federal

(b) Client

2. Level of client
- personal
- enterprise

(c) Service

3. Type of service
- obstruction of justice
- lobbyism
- extortion for rendering official services
- improper categorization
- providing security (cover-up)
- obstruction of competition

4. Frequency of the service
- once
- multiple

(d) Interaction

5. Initiator
- client
- official

6. Bribe size
- $10,000 or less
- from $10,001 to $100,000
- $100,000 or more

7. Type of corruption
- infringement
- extortion

8. Corruption network
- isolated actor
- network within corrupt office
- between-office network
- client network

(e) Environment

9. Condition of corruption
- regular order of action
- inspection and monitoring
- fuzziness of regulations
- irregular event

10. Branch
- government
- low-enforcement
- other

11. Punishment
- none
- administrative
- arrest followed by release
- arrest with unknown consequences
- arrest with (potential) imprisonment

Seven of the variables:
2. Level of client
4. Frequency
5. Initiator
6. Bribe size
7. Type of corruption
8. Corruption network
11. Punishment

will be considered ranked (in the presented order of categories), in fact, quantitative with the ranks used as numerical scores. The others are nominal, with each of their categories coded as a binary yes/no variable. This produces altogether 24 features that are then treated as constituting the variable space.

From the initial set of 68 articles, 55 items present clear-cut cases which can be coded, more or less unanimously, by the features above.

The problem to be addressed with clustering: whether any patterns of corruption exist in the data?

2.2. Data processing

According to the prescriptions above, the data processing includes the following steps:

1. **Data standardisation.** This is done by subtracting the feature averages (grand means) from all entries and by follow-up dividing them by the feature ranges. For a binary feature corresponding to a qualitative category, this reduces to subtraction of the category proportion, p, from all the entries that become this way either $1-p$, for "yes", and $-p$, for "no".

2. **Performing Separate/Conquer.** Application of Separate/Conquer to the pre-processed data matrix with the reference point in the space origin, 0, has produced 13 clusters presented in Table 1. They explain 64 % of the data dispersion.

3. **Initial setting for K-Means.** There are only 5 clusters that have more than three elements according to Table 1. This defines the number of clusters as well as the initial setting: first elements of the five larger clusters, indexed as 5, 12, 4, 1, and 11, are taken as the initial centroids.

4. **Performing K-Means.** K-Means applied to the data has produced five clusters presented in Table 2. They explain 45 % of the data scatter. Thus reduced proportion of the explained data can be explained by the reduced number of clusters.

Table 1

Characteristics of clusters found by Separate/Conquer algorithm in data

Cluster	#	Elements	Contribution, %
1	7	5,16,23,27,28,41,42	9.8
2	1	25	2.2
3	2	17,22	3.3
4	1	49	2.2
5	1	2	2.1
6	1	35	2.1
7	13	12,13,20,33,34,38,39,43, 45,47,48,50,51	10.7
8	9	4,6,9,10,21,26,30,31,40	10.2
9	5	1,3,15,29,32	6.3
10	2	7,52	3.3
11	3	8,14,36	3.4
12	8	11,24,37,44,46,53,54,55	7.8
13	2	18,19	2.6

2.3 Interpretation of the clusters

Since individual corruption cases are not of interest here, only two levels of interpretation will be presented.

Let us look at cluster 1. Its most contributing features are: Other branch (877%), Improper

Table 2

Characteristics of clusters found with K-Means algorithm in data

Cluster	#	Elements	Contribution, %
1	8	5,16,23,25,27,28,41,42	10.0
2	19	7,8,12,13,14,20,33,34,35,3638,39, 43,45,47,48,50,51,52	9.8
3	10	4,6,9,10,21,22,26,30,31,40	10.0
4	7	1,3,15,17,29,32,49	7.0
5	11	2,11,18,19,24,37,44,46,53, 54,55	8.1

categorisation (439 %), and Level of client (242 %). Here and further the values in parentheses are ratios of feature contributions to clusters and of those to the data scatter according to formula (*). By looking at the cluster centroid, one can find specifics of the features in the cluster. In particular, all its cases fall in branch «Other» comprising such bodies as universities or hospitals. In each of the cases the issue was of a personal matter, and most time (six of the eight cases) the service provided was based on re-categorisation of the client into a better category. As the three features are most contributing to the cluster, one can be sure that the tendencies expressed are relevant to the cluster as opposed to the rest of the data set. Indeed, the category Other branch (of

variable number 10) appears to be distinctively describing the cluster: there are eight cases in this category and all of them belong to the cluster.

There are nineteen cases in cluster 2, and most salient features are: Obstruction of justice (467 %), Law enforcement (379 %), and Occasional event (251 %). The centroid values of these features show that all corruption cases in this cluster have occurred in the law enforcement system. They are mostly done via obstruction of justice for occasional events. The fact that all cluster cases occurred in a particular branch, the law enforcement system, is not sufficient for distinctively describing the cluster since law enforcement corruption relates to 62 % of all the 55 cases, not just nineteen. Two more conditions have been found by algorithm APPCOD to make the description distinctive: the cases occurred at office levels higher than Organisation and the cases did not involve cover-up.

Cluster 3 contains ten cases for which most salient features are: Extortion in variable 3 (474 %), Organisation (289 %), and Government (275 %). Nine of the cases occurred in the Government system overwhelmingly at the level of organisation and, also overwhelmingly, the office workers extorted money for rendering services that they were obliged to do anyway. Also, the client level here is always of an organisation, though this feature is not that salient as the other three. This is the cluster tendency. However, to put it at the level of distinctive description appears to be a complicated job, because no conjunction of single categories can do this. Algorithm APPCOD found a distinctive description involving two arithmetically combined variables, var1=Extortion-Obstruction of justice and var2 = Extortion + Bribe. The entire description consists of four conjunctive terms stating that $1 \geq var1 \geq 0$ and $3 \geq var2 \geq 2$ and neither Lobbyism/Protection nor Inspection occurred. The inequalities show the real meaning of the combined variables. The first inequality, in fact, states that in the cases in which no Extortion occurred, Obstruction of justice also was not on the agenda. The second inequality states the condition that under Extortion for rendering legal services the level of Bribe was smaller (1 or 2) than in the cases when Extortion did not apply (2 or 3). All these, basically, say that the cluster pertains to Extortion and some subtler conditions to get a distinctive description.

Cluster 4 contains seven cases, and its salient features are: Lobbyism/Protection (913 %), Government (391 %), and Federal (338 %). All the cases occurred in the government legislative and executive branches. The service provided was mostly Lobbyism/Protection (six of seven cases). Federal level of corrupt office was not that frequent, two cases only, but it was unexpectedly frequent as the two cases are just half of the total number of cases, four, when Federal level of office was involved. Algorithm APPCOD found some more features to distinctively describe the cluster with a conjunction of four statements: the cases involve Government and Infringement as Type of corruption (variable 7); they occur at Federal and other levels higher than Organisation and they never involve Cover-up as the corruption service.

Cluster 5 consists of eleven cases and pertains to two salient features: Cover-up (807 %) and Inspection (469 %). All of them involve Cover-up as the service provided, mostly in inspection and monitoring activities (nine cases). A distinctive description of the cluster found by APPCOD conjuncts two statements: it is always Cover-up but not at the level of Organisation.

Overall, the cluster structure shows that Branch is the defining variable in Russian corruption looked through the media glass. Different branches tend to involve

different corruption services. The government corruption involves either Extortion for rendering their free services to organisations (Cluster 3) or Lobbyism/Protection (Cluster 4). The law enforcement corruption in higher offices makes it for either Obstruction of justice (Cluster 2) or Cover-up (Cluster 5). Actually, Cover-up does not exclusively belong in the law enforcement branch: in fact, it relates to offices that are to inspect and monitor commercial and other activities (Cluster 5). Corruption cases in branch Other involve re-categorisation of individual cases into better suitable categories.

Literature

1. Mirkin B. Concept learning and feature selection based on square-error clustering // Machine Learning. – 1999. – № 35. – С. 25-40.
2. Mirkin B. Reinterpreting the category utility function // Machine Learning. – 2001. – № 45. – С. 219-228.
3. Mirkin B., Ritter O. A feature-based approach to discrimination and prediction of protein folding // Genomics and Proteomics / Edit. by S. Suhai . – New York: Kluwer Academic / Plenum Publishers. – С. 157-177.

Материал поступил в редакцию 09.04.02.