Barometr Regionalny
Tom 12 nr 3

Numerical Data Clustering Algorithms in Mining Real Estate Listings

Krzysztof Pancerz
University of Management and Administration in Zamość, Poland
University of Information Technology and Management in Rzeszów, Poland

Olga Mich
University of Management and Administration in Zamość, Poland

Abstract

In the paper, we propose a method for mining real-estate listings using clustering algorithms intended for numerical data. The presented approach is based on information systems over ontological graphs. Such information systems have been proposed to deal with data in the form of concepts linked by different semantic relations. A special attention is focused on preprocessing steps transforming advertisements in the textual form into information systems defined over ontological graphs, as well as on encoding attribute values for clustering algorithms.

Keywords: clustering, data mining, information systems, ontological graphs, real-estate listings

Introduction

In general, in data mining, we have a data set of objects (also called cases, examples, or instances), each of which comprises the values of a number of variables (often called attributes in data mining). There are two types of data, which are treated in different ways:

- labeled data — data with a specially designated attribute, and
- unlabeled data — data without any specially designated attribute.

Data mining of unlabeled data is known as unsupervised learning. Clustering is one of the unsupervised learning techniques used in data mining. It is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other groups (Bramer 2007; Cios et al. 2007). In economics, financial and marketing applications (e.g., those connected with researching regional markets), there are obvious benefits to be derived from grouping together similar objects, for example, finding regions whose economies are similar, finding companies that have similar financial performance, finding customers with similar behavior, etc. In the paper, we are interested in mining real-estate listings to find groups of similar advertisements. Such a problem may appear in intelligent guidance and suggestions for customers purchasing real estates. Intelligent guidance and suggestion systems implementing analysis of purchasing patterns of users play an increasingly large role in marketing (Rosenfeld 2011).

Real estate listings include a special kind of data. Such data are stored in the textual form. Each advertisement consists of loosely coupled words (terms, concepts) rather than full, grammatically correct sentences. Moreover, underlying data are of qualitative character. One of challenges posed by such data is understanding data semantics and semantic relations between them which can be treated as some external knowledge useful in data mining processes. Moreover, a lot of data mining methods work with numerical data and, for such methods, textual data have to be encoded. In the paper, we propose a method of mining real-estate listings using clustering algorithms intended for numerical data. The presented approach is based on information systems

over ontological graphs described briefly in Section 1. In the proposed method, we can distinguish two main parts. The first part comprises preprocessing steps transforming advertisements in the textual form into information systems defined over ontological graphs and next into information systems with numerical attribute values. The second part includes a proper clustering procedure. The whole process is described in Section 2. In Section 3, we show a practical example explaining the proposed approach. Some conclusions, as well as the main direction for further work, are given in Section 4.

1 Basics

In (Pawlak 1991), information (decision) systems were proposed as the knowledge representation systems. In simple case, they consist of vectors of numbers or symbols (attribute values) describing objects from a given universe of discourse. Formally, an information system IS is a quadruple $(U,A,\mathbf{V},f)$ consisting of:

U — a nonempty, finite set of objects (cases, examples, instances),
A — a nonempty, finite set of attributes (features, properties),
$\mathbf{V}$ — a family of attribute value sets (with each attribute from A, a set of its values is associated),
f — an information function assigning, to each attribute-object pair (a,u), one value from the set of values associated with the attribute a.

It is worth noting that a given definition of an information system is a basic one. In the literature, other information systems are considered, for example, multi-valued information systems, information systems with missing values (incomplete information systems), etc. A brief account of different information systems is given, among others, in (Qiu et al. 2012). Any information system can be presented in a tabular form. Such a form is called an information table. In information tables, rows represent objects whereas columns correspond to attributes. Entries of the tables (intersections of rows and columns) are attribute values.

In many applications, we are interested in mining textual data in the form of concepts (words, terms). Therefore, in (Pancerz 2012), ontologies were incorporated into information systems — i.e., attribute values were considered in the ontological (semantic) space. Ontologies deliver the knowledge about semantic relations between concepts (Neches et al. 1991). Formally, the ontology can be represented by means of graph structures, called here ontological graphs. For a given ontology, an ontological graph $\mathbf{OG} = (C,E,R,\rho)$ consists of:

C — a nonempty, finite set of nodes representing concepts in the ontology,
E — a finite set of edges representing relations between concepts from C,
R — a family of semantic descriptions of types of relations (represented by edges) between concepts,
ρ — a function assigning a semantic description of the relation to each edge.

Exemplary ontological graphs representing the real-estate domain are shown in Section 3.

In (Murphy 2008), two general types of relations between words (concepts) were distinguished:

- paradigmatic relations which are relations between words belonging to the same grammatical category
- syntagmatic relations which are relations between words that go together in a syntactic structure

It is worth noting that paradigmatically related words are, to some degree, grammatically substitutable for each other. In the presented approach, we are interested in paradigmatic relations because, in real-estate listings, we do not take into consideration a semantic structure of sentences, but we use some knowledge about concepts (terms) included in them, for example, whether they are synonyms, whether one concept can be replaced with another, for example, more general one, etc. Paradigmatic relations are often called semantic relations or lexical relations. There are a lot of semantic relations defined in the literature. For example, in (Chaffin, Herrmann, and Winstonc 1988), the authors provided a list of 31 semantic relations that are broken into different categories. In some of our previous research (Pancerz and Lewicki 2014; Pancerz and Mich 2014), we have taken into consideration, only three types of semantic relations:

- synonymy
- antonymy
- hyponymy/hyperonymy

Synonymy concerns concepts with a meaning that is the same as, or very similar to, another concepts. Antonymy concerns concepts which have the opposite meaning to the other ones. Hyponymy/hyperonymy determines narrower/broader meaning. Hyponymy concerns more specific concepts than the other ones. Hyperonymy concerns more general concepts than the other ones. In ontological graphs, these relations will be denoted as follows:

- *isSyn* denotes synonymy (if two concepts u and v are in the *isSyn* relation, it means that u is a synonym of v and vice versa)
- *isAnt* denotes antonymy (if two concepts u and v are in the *isAnt* relation, it means that u is a antonym of v and vice versa)
- *isGen* denotes hyponymy (if two concepts u and v are in the *isGen* relation, it means that u is a hyponym of v or u is generalized by v)
- *isSpec* denotes hyperonymy (if two concepts u and v are in the *isSpec* relation, it means that u is a hyperonym of v or u is specialized by v)

Additionally, we take into consideration a semantic relation called "being an instance". Being an instance concerns an example (instance) of a given concept. This kind of relations is important in mining real-estate listings because they include, for example, instances of places.

In the presented approach, we use simple information systems over ontological graphs defined in (Pancerz 2012). In such systems, attribute values are concepts from ontologies assigned to attributes. Formally, a simple information system over ontological graphs *SIS* is a quadruple $(U,A,\mathbf{OG},f)$ consisting of:

U — a nonempty, finite set of objects,
A — a nonempty, finite set of attributes,
$\mathbf{OG}$ — a family of ontological graphs associated with attributes (with each attribute from A, an ontological graph is associated),
f — an information function assigning, to each attribute-object pair (a,u), one concept from the ontological graph associated with the attribute a.

An exemplary information system over ontological graphs, representing the real-estate domain, is shown in Section 3.

2 Procedure

In this section, we show the procedure of clustering advertisements included in real-estate listings using algorithms intended for numerical data. A starting point for various clustering algorithms is an information system (information table) consisting of descriptions of objects in the form of numerical vectors. This is dictated by the fact that the majority of standard similarity (distance) measures, including the most known Euclidean distance, used to compare objects in such algorithms, is defined for numerical vectors (Gan, Ma, and Wu 2007). The proposed procedure can be divided into two main parts. In the first part, called the preprocessing part, advertisements in the textual form are transformed into information systems defined over ontological graphs and next into encoded information systems including numerical vectors describing objects (advertisements). The procedure transforming advertisements in the textual form into information systems over ontological graphs has been proposed by us in (Pancerz and Mich 2014). To obtain an encoded information system, we can use the idea proposed in (Pancerz and Lewicki 2014). In the second part, a clustering procedure is performed. The scheme of the whole procedure is presented in figure 1.

As one can see in figure 1, the following steps can be distinguished:

- stemming — defining basic grammatical forms (roots) for particular words existing in advertisements, for example, using a quite popular Porter stemming algorithm (Porter 1980)
- attributation — assigning concepts (built from words) existing in advertisements to proper attributes as their values, according to defined ontological graphs

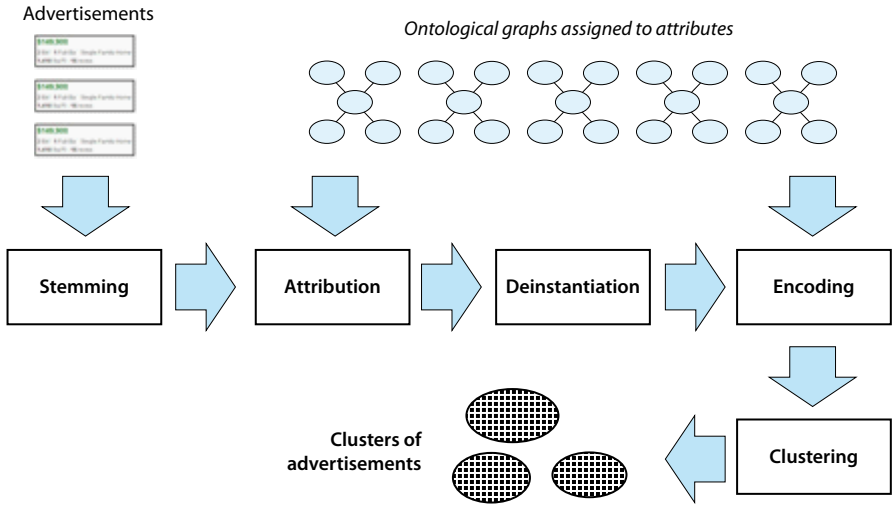

Fig. 1. A procedure for clustering advertisements included in real-estate listings using algorithms intended for numerical data

- deinstantiation — replacing instances existing in advertisements with the most specific concepts (with respect to the hyponymy/hyperonymy relation) whose instances they are
- encoding — transforming vectors of concepts describing advertisements into proper vectors of numbers using the additional knowledge about semantic relationships between concepts from ontological graphs assigned to attributes
- clustering — creating groups of encoded numerical vectors representing advertisements, in such a way that vectors in one group are very similar (with respect to a given similarity measure) and vectors in different groups are quite distinct

It is worth noting that deinstantiation is an important step if we are interested in a more general knowledge derived from real-estate listings, for example, some client is interested only in houses in a village (not a particular one). The presented procedure is explained on the example given in Section 3.

3 Example

In this section, we explain the procedure described in Section 2 on the example of the real-estate listings. We have collected one hundred real estate advertisements which have been transformed into an information system over ontological graphs shown partially in table 1. After performing our procedure presented in Section 2, each advertisement becomes one object (row) in an information system. Each object (advertisement) is described by three attributes (Pancerz and Lewicki 2014):

- offer type — with attribute values (concepts) included in the ontological graph shown in figure 2
- property type — with attribute values (concepts) included in the ontological graph shown in figure 3
- preferred place — with attribute values (concepts) included in the ontological graph shown in figure 4

As one can see in the ontological graphs associated with attributes, concepts are in some hyponymy/hyperonymy or antonymy relations.

To encode symbolic values (concepts) appearing in a simple information system over ontological graphs we can use the method proposed in (Pancerz and Lewicki 2014). In that method, the additional knowledge about semantic relations between concepts from ontological graphs assigned to attributes is utilized. For each attribute a and each object u in a simple information system over ontological graphs, we define a fuzzy characteristic function assigning a value from the interval [0,1] to each concept c from the ontological graph associated with a in the following way:

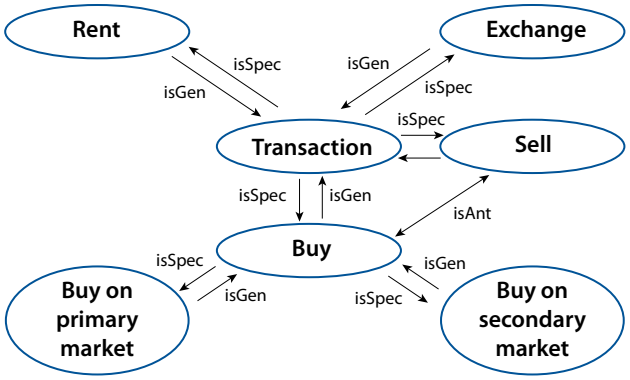

Fig. 2. The ontological graph associated with the attribute Offer type

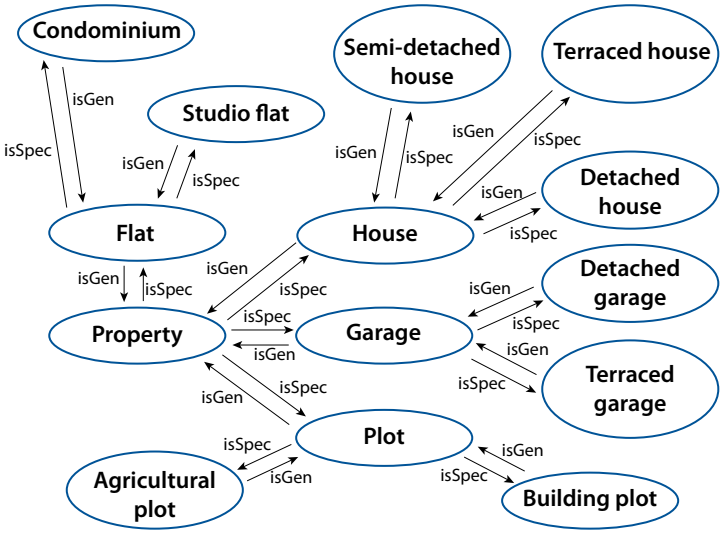

Fig. 3. The ontological graph associated with the attribute Property type

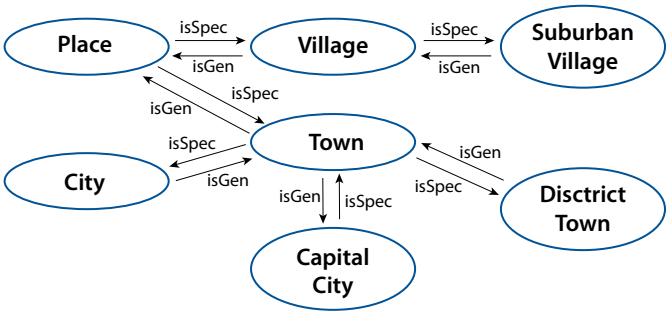

Fig. 4. The ontological graph associated with the attribute Preferred place

Tab. 1. A fragment of an information system over ontological graphs describing advertisements

ID	Offer type	Property type	Preferred place
1	Sell	House	City
2	Sell	House	City
3	Sell	House	Capital city
4	Sell	House	Suburban village
5	Rent	Flat	Capital city
6	Rent	Flat	Capital city
...	...	...	...

- 1 if the value v of the attribute a for the object u is the concept c or v is a synonym of c or v is a hyperonym (direct or indirect) of v
- $e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}$ if the value v of the attribute a for the object u is a hyponym of the concept c or v and c have a common hyperonym (direct or indirect), excluding situation when v and c are antonyms, where l is the length of the shortest path between v and c in the ontological graph, h is the depth of the subsumer in the hierarchy, α is a constant and β is a smoothing factor (see Li et al. 2003)
- 0 otherwise

Value 1 means that the concept c is semantically related to the concept v describing the object u on the attribute a according to two relations: synonymy or hyponymy/hyperonymy. This choice of relations seems to be reasonable because if a given concept v describes the object u, then all its synonyms can be treated as v as well as all hyponyms of v can also be treated as v according to the "is-a" property of a generalization (Brachman 1983).

It is easy to see that the encoded information system includes as many attributes as many concepts there are in all ontological graphs associated with attributes. A fragment of the encoded information system, corresponding to the fragment of the information system over ontological graphs from table 1, is shown in table 2.

Having the encoded information system shown partially in table 2, we can use any clustering algorithm working with numerical data. The reader can find an overview of a variety of algorithms, for example, in (Gan, Ma, and Wu 2007). In our example, we have used an agglomerative hierarchical clustering algorithm called *agnes*, described in (Kaufman and Rousseeuw 1990) and

Tab. 2. A fragment of the encoded information system describing advertisements

Sell	Rent	House	Semi-detached house	Terraced house	...	City	Capital city	...
1	0	1	1	1	...	1	1	...
1	0	1	1	1	...	1	1	...
1	0	1	1	1	...	0,44	1	...
1	0	1	1	1	...	0,37	0,37	...
0	1	0,44	0,29	0,29	...	0,44	1	...
0	1	0,44	0,29	0,29	...	0,44	1	...
...	...	...	...	...	...	...	...	...

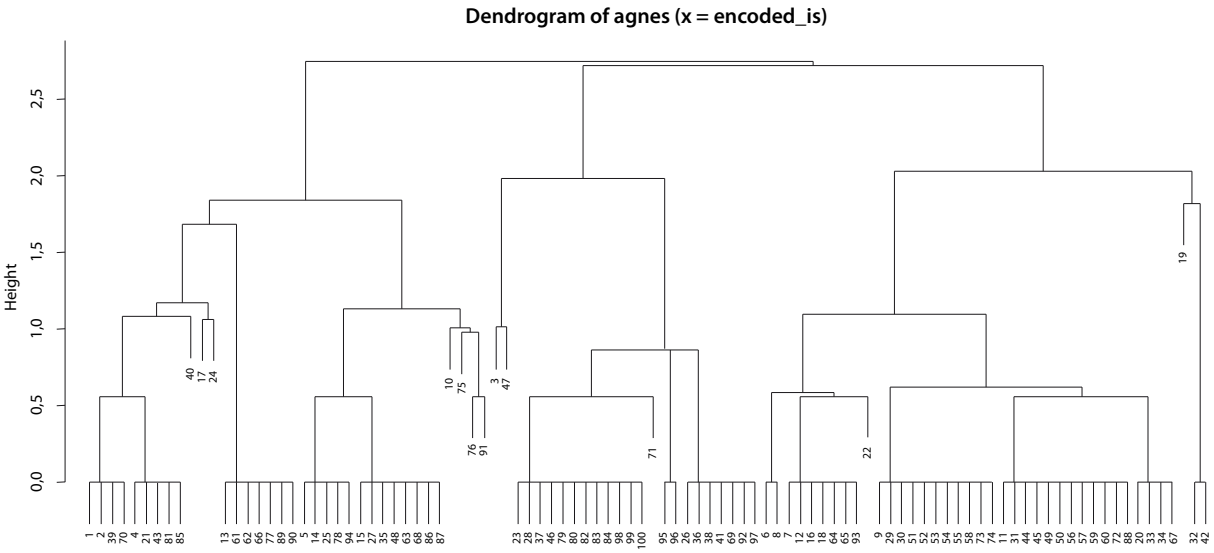

Fig. 5. The dendrogram obtained after clustering advertisements using the agnes algorithm implemented in the R environment

implemented in the R environment[1]. In the agglomerative clustering algorithm, we start with each object in a cluster of its own and repeatedly merge the closest pair of clusters (with respect to a given similarity measure) until we end up with one cluster containing all objects (Bramer 2007). The result of the agglomerative clustering algorithm is presented in the form of the so-called dendrogram visualizing consecutive steps of merging clusters. In Figure 5, the result, obtained in the R environment, of agglomerative clustering of our 100 advertisements, is shown in the form of a dendrogram.

Information about similar advertisements (in the semantic space) can be helpful to suggest (recommend) real-estates for purchasing by some clients.

Conclusions and Further Work

In the paper, we have shown how to transform advertisements included in real-estate listings to find groups of similar ones using clustering algorithms intended for numerical value. An analogous procedure can be performed for other data mining techniques, for example, classification, knowledge discovery, etc. (Pancerz and Lewicki 2014; Pancerz and Mich 2014). The main direction for further work is to define similarity measures enabling us to use clustering algorithms directly for information systems over ontological graphs — i.e., without the necessity of encoding symbolic (concept) descriptions of objects (advertisements).

References

Brachman, R.J. 1983. "What Is-a Is and Isnt — an Analysis of Taxonomic Links in Semantic Networks." *Computer* no. 16 (10):30–36.

Bramer, M.A. 2007. *Principles of Data Mining, Undergraduate Topics in Computer Science*. London: Springer.

Chaffin, R., D.J. Herrmann, and M. Winston. 1988. "An Empirical Taxonomy of Part-Whole Relations. Effects of Part-Whole Relation Type on Relation Identification." *Language, Cognition and Neuroscience* no. 1 (3):17–48.

Cios, K.J., W. Pedrycz, R.W. Swiniarski, and L. Kurgan. 2007. *Data Mining. A Knowledge Discovery Approach*. New York: Springer.

Gan, G., C. Ma, and J. Wu. 2007. *Data Clustering. Theory, Algorithms, and Aplications, ASA-SIAM series on statistics and applied probability*. Philadelphia, Pa.; Alexandria, Va.: SIAM; American Statistical Association.

Kaufman, L., and P.J. Rousseeuw. 1990. *Finding Groups in Data. An Introduction to Cluster Analysis, Wiley series in probability and mathematical statistics, Applied probability and statistics*. New York: Wiley.

Li, Y.H., Z.A. Bandar, and D. McLean. 2003. "An Approach for Measuring Semantic Similarity Between Words Using Multiple Information Sources." *IEEE Transactions on Knowledge and Data Engineering* no. 15 (4):871–882.

Murphy, M.L. 2008. *Semantic Relations and the Lexicon. Antonymy, Synonymy, and Other Paradigms*. Cambridge, UK; New York: Cambridge University Press.

Neches, R., R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W.R. Swartout. 1991. "Enabling Technology for Knowledge Sharing." *AI Magazine* no. 12 (3):36–56.

Pancerz, K. 2012. "Toward Information Systems over Ontological Graphs." In *Rough Sets and Current Trends in Computing*, edited by J. Yao, Y. Yang, R. Słowiński, S. Greco, H. Li, S. Mitra and L. Polkowski, 243–248. Berlin–Heidelberg: Springer.

Pancerz, K., and A. Lewicki. 2014. "Encoding Symbolic Features in Simple Decision Systems over Ontological Graphs for PSO and Neural Network Based Classifiers." *Neurocomputing* no. 144:338–345. doi: 10.1016/j.neucom.2014.04.038.

Pancerz, K., and O. Mich. 2014. Mining Real-Estate Listings Based on Decision Systems over Ontological Graphs: Extended Abstract. Paper read at Proceedings of the Workshop on Concurrency, Specification and Programming (CS&P'2014), 2014.09.29–10.01, at Chemnitz, Germany.

1. See: http://www.r-project.org/.

Pawlak, Z. 1991. *Rough Sets. Theoretical Aspects of Reasoning about Data, Theory and decision library Series D, System theory, knowledge engineering, and problem solving.* Dordrecht-Boston: Kluwer Academic Publishers.

Qiu, T., L. Liu, L. Duan, S. Zhou, and H. Huang. 2012. "A Rough Set Model for Incomplete and Multi-Valued Information Systems." *International Journal of Digital Content Technology and its Applications* no. 6 (20):53–61. doi: 10.4156/jdcta.vol6.issue20.6.

Rosenfeld, L. 2011. *Search Analytics for Your Site. Conversations with Your Customers.* Brooklyn, N.Y.: Rosenfeld Media.