Mining Spatial Rules by Finding Empty Intervals in Data

Alexandr Savinov

Fraunhofer Institute for Autonomous Intelligent Systems
Schloss Birlinghoven, Sankt-Augustin, D-53754 Germany
savinov@ais.fraunhofer.de
http://www.ais.fhg.de/~savinov/

Abstract. Most rule induction algorithms including those for association rule mining use high support as one of the main measures of interestingness. In this paper we follow an opposite approach and describe an algorithm, called Optimist, which finds all largest empty intervals in data and then transforms then into the form of multiple-valued rules. It is demonstrated how this algorithm can be applied to mining spatial rules where data involves both geographic and thematic properties. Data preparation (spatial feature generation), data analysis and knowledge postprocessing stages were implemented in the SPIN! spatial data mining system where this algorithm is one of its components.

1 Introduction

The conventional approach to rule induction consists in finding anomalies by searching for intervals with surprisingly high values of probability distribution representing the data semantics and the larger such an interval the better. For instance, in association rule mining patterns are generated in the form of itemsets and their interestingness is measured by support (the number of objects satisfying both condition and conclusion) and confidence (the number of objects satisfying the rule consequent among those satisfying the antecedent). For example, we might infer a rule where some item, e.g., high long-term illness, under some conditions has 99% confidence. Implicitly it means that other (mutually exclusive) items, e.g., medium and low long-term illness have much less probability very close to 0. Thus the rule semantics can be reformulated as incompatibility of some target values with items in condition. Interesting rules then can be generated by finding item combinations that never occur in data. The goal is still finding some kind of anomalous behaviour but the main distinction from traditional approaches is that we are trying to find empty areas instead of high frequency areas in the data. A related approach to mining association rules based on this principle is described in [4-7] were empty intervals among numeric attributes are called holes in data. The holes are found by using the algorithm [8-9] from computational geometry. In this paper we apply an original rule induction algorithm, called Optimist [1-3], which works with finite value attributes and generates rules for one pass through the data set by using a method of sectioned vectors.

It is estimated that 80% of data are geo-referenced and recently spatial data mining area has been paid significant attention. Particularly, spatial rule induction offers great potential benefits for solving the problem of spatial intelligent data analysis. In this paper we describe how the rule induction algorithm based on finding largest empty intervals in data can be applied to spatial data analysis. Since analysis itself is known to take only a small potion of the whole knowledge discovery process while such tasks as data preparation and postprocessing take most of time we integrated our rule induction algorithm into the SPIN! spatial analysis system. The SPIN! system integrates several data mining methods adapted to the analysis of spatial data, e.g., multi-relational subgroup discovery and spatial cluster analysis, and combines them with thematic mapping functionality for visual data exploration, thus offering an integrated environment for spatial data analysis [10-11].

2 Generating Largest Empty Intervals

Attributes $x_1, x_2, \ldots, x_n$ are assumed to take a finite number of *values,* n_i from their domains $A_i = \{a_{i1}, a_{i2}, \ldots, a_{in_i}\}$. All combinations of values $\omega = \langle x_1, x_2, \ldots, x_n \rangle \in \Omega = A_1 \times A_2 \times \ldots \times A_n$ form the *state space* or *universe of discourse.* Each record from a data set corresponds to one combination of attribute values or a point. If for a combination of values a record in the data set exists then it is said to be possible. Otherwise the point is impossible. To represent the data semantics as Boolean distribution over the universe of discourse we use the method of sectioned vectors and matrices [2,3]. The idea of the method is that one vector can represent a multidimensional interval of possible or impossible points (called also positive and negative internal, respectively). Each vector consists of 0s and 1s, which are grouped into sections separated by dots and corresponding to all attributes. A section consists of n_i components corresponding to all attribute values. For example, 01.010.0101 is a sectioned vector for three attributes taking 2, 3 and 4 values. A sectioned vector associates *n* components with each point from Ω (one from each section). The position of these components in the vector corresponds to the point coordinates. To represent negative intervals we use disjunctive interpretation of sectioned vector. It means that the point is assigned 0 if all its components in the vector are 0s, and it is assigned 1 if at least one the components is 1. For example, the point $\langle a_{11}, a_{21}, a_{31} \rangle$ is impossible according to the above vector semantics while the point $\langle a_{11}, a_{22}, a_{33} \rangle$ is possible since the component corresponding to a_{22} is equal to 1.

The idea of the algorithm for finding largest empty intervals consists in representing data semantics by a set of negative sectioned vectors and updating it for each record. Initially the data is represented by the only empty interval consisting of all 0s and making all points impossible. After the first record is added it is split into several smaller negative intervals so that the point corresponding to this record becomes possible. For example, addition of the record 0$\underline{1}$.00$\underline{1}$.000$\underline{1}$ (where 1s correspond to its values) to the interval 00.010.0100 splits it into three new intervals: 0$\underline{1}$.010.0100, 00.01$\underline{1}$.0100, and 00.010.010$\underline{1}$ (changed components are underlined). During this procedure very small intervals with a lot of 1s are removed since they generate very specific rules and only the top set of the largest intervals is left.

Once largest empty intervals have been found they can be easily transformed into rules by negating sections , which should be in antecedent. For example, the vector $\{0,1\} \vee \{0,1,0\} \vee \{0,1,0,1\}$ can be transformed into the implication $\{1,0\} \wedge \{1,0,1\} \rightarrow \{0,1,0,1\}$ interpreted as the rule IF $x_1 = \{a_{11}\}$ AND $x_2 = \{a_{21}, a_{23}\}$ THEN $x_3 = \{a_{32}, a_{34}\}$. The rules are filled in by statistical information in the form of the target value frequencies within the rule condition interval (for one additional pass through the data set). In other words, each value in conclusion is assigned its frequency within the condition interval, e.g., IF $x_1 = \{a_{11}\}$ AND $x_2 = \{a_{21}, a_{23}\}$ THEN $x_3 = \{a_{32} : 145, a_{34} : 178\}$, which is obviously more expressive. Here 145 means that the value a_{32} occurs 145 times within the selected interval.

3 Mining Interesting Spatial Rules

The Optimist algorithm has been implemented as one of SPIN! spatial data mining system components [10-11] (Fig. 1). It is tuned by a set of algorithm parameters such as maximal number of patterns (empty intervals) and execution on the client or on the server. Input data for the algorithm is specified by a standard SPIN! query component, which uses a separate connection component to access a database. The spatial rules generated by the algorithm are stored in rule base component. When appropriately connected this minimal set of components implements the conventional knowledge discovery cycle. The analysis starts from specifying database and query, which can produce data in the necessary format. In our case we need data columns to take only a finite number of values. Since most of source data had continuous attributes we applied SPIN! optimal discretization algorithm [12]. Once columns have been discretized it is necessary to generate spatial attributes. For this purpose we used spatial functionality of Oracle 9i database where objects are represented by means of special built-in geometry type. Using such a representation a query can combine spatial information with thematic data describing objects located in space. It is important that various spatial properties can be automatically generated by the database with the help of spatial predicates and relationships.

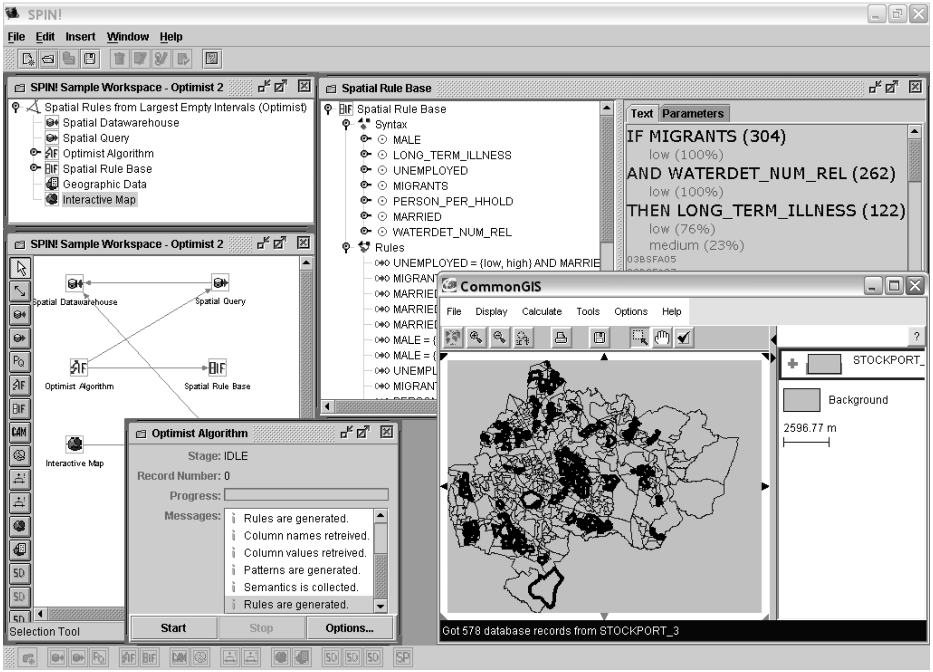
We used UK 1991 census data for Stockport[1], one of the ten districts in Greater Manchester, UK. The analysis was carried out at the level of enumeration districts (the lowest level of aggregation) characterized by such attributes as person per household, cars per household, migration, long-term illness, unemployment and other census statistics. Spatial information was available as coordinates and borders of such objects as enumeration districts, water, roads, streets, railways, bus stops etc. For typical analysis we might be interested in finding dependencies among different thematic and spatial attributes, for example, what spatial and non-spatial factors influence long-term illness. As spatial characteristic we define an attribute, which counts

[1] All data are provided by the Manchester University and Manchester Metropolitan University.

the number of water resources belonging to each enumeration district calculated by means of SQL statement with spatial join. The final result set produced by SQL query is a normal table, which can be directly analyzed by the Optimist rule induction algorithm. Here is an examples, which has been generated by such an analysis where MARRIED is the percentage of married people and WATER_NUM_REL is a characteristic of water resources in the enumeration district:

IF MARRIED (461) {high (46%) OR medium (53%)}
AND WATER_NUM_REL (447) {low (58%) OR medium (41%)}
THEN LONG_TERN_ILLNESS (358) {low (68%) OR medium (31%)}

The produced rules can be shown in its own window where the rules can be studied in details. However, the SPIN! system provides much more powerful method by using linked displays and interactive visualisation functionality [13]. The idea is that objects described in one view can be simultaneously visualised in other views. In our case the rules describe enumeration districts while these very districts can be simultaneously shown on the map. Moreover, as we select some rule all objects, which satisfy its left hand side are dynamically highlighted on the map so that we can easily see how they are spatially distributed (Fig. 1). For example, we might find that enumeration districts satisfying some rule and thus having interesting characteristics in terms of the target attribute form a cluster or have more complex spatial configuration, e.g., with respect to other geographic objects such roads and cities.



IF MIGRANTS (304)
low (100%)
AND WATERDET_NUM_REL (262)
low (100%)
THEN LONG_TERM_ILLNESS (122)
low (76%)
medium (23%)
03BSFA05

Stage: IDLE
Record Number: 0
Progress:
Messages:
Rules are generated.
Column names retreived.
Column values retreived.
Patterns are generated.
Semantics is collected.
Rules are generated.

Got 578 database records from STOCKPORT_3

Fig. 1. Visualization of spatial rules simultaneously and interactively with the map and other views in the SPIN! system. As one rule is selected in the upper right view all objects satisfying its condition are dynamically highlighted on the map in the lower right window.

4 Conclusion

In the paper we described an approach to mining spatial rules by finding largest empty intervals in multidimensional space. The advantage of the algorithm is that it directly generates highly expressive multiple-valued rules for one pass over the data set (additional pass is needed for generating additional rule statistics). Particularly it does not require the data set to be in memory and hence can be applied to very large tables. Combined with additional data preprocessing and geographic visualisation components within SPIN! spatial data mining system it allows for carrying out complex analysis of real world data involving both spatial and non-spatial attributes.

References

1. A.A. Savinov, Mining possibilistic set-valued rules by generating prime disjunctions, Proc. 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases — PKDD'99, Prague, Czech Republic, September 15-18, 1999, 536-541.
2. A.A. Savinov. Application of multi-dimensional fuzzy analysis to decision making. In: Advances in Soft Computing — Engineering Design and Manufacturing, R. Roy, T. Furuhashi and P.K. Chawdhry (eds.), Springer-Verlag, London, 1999.
3. A.A. Savinov, An algorithm for induction of possibilistic set-valued rules by finding prime disjunctions, In: Soft computing in industrial applications, Suzuki, Y., Ovaska, S.J., Furuhashi, T., Roy, R., Dote, Y. (Eds.), Springer-Verlag, London, 2000.
4. Bing Liu, Liang-Ping Ku and Wynne Hsu, "Discovering Interesting Holes in Data," *Proceedings of Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97)*, pp. 930-935, August 23-29, 1997, Nagoya, Japan.
5. Bing Liu, Ke Wang, Lai-Fun Mun and Xin-Zhi Qi, "Using Decision Tree Induction for Discovering Holes in Data," *Pacific Rim International Conference on Artificial Intelligence (PRICAI-98)*, 182-193, 1998
6. Liang-Ping Ku, Bing Liu and Wynne Hsu, "Discovering Large Empty Maximal Hyperrectangles in Multi-dimensional Space," *Technical Report*, Department of Information Systems and Computer Science (DCOMP), National University of Singapore, 1997.
7. Jeff Edmonds, Jarek Gryz, Dongming Liang, and Renée J. Miller, *Mining for Empty* Rectangles *in Large Data Sets*. In Proceedings of the 8th International Conference on Database Theory (ICDT), London, UK, January 2001, pp. 174-188.
8. M. Orlowski. A New Algorithm for the Largest Empty Rectangle Problem. Algorithmica, 5(1):65--73, 1990.
9. B. Chazelle, R. L. Drysdale, and D. T. Lee. Computing the largest empty rectangle. SIAM J. Comput., 15:300-315, 1986.
10. M. May, A. Savinov, An Architecture for the SPIN! Spatial Data Mining Platform, Proc. New Techniques and Technologies for Statistics, NTTS 2001, 467-472, Eurostat, 2001
11. M. May, A. Savinov, An integrated platform for spatial data mining and interactive visual analysis, Data Mining 2002, 25-27 September 2002, Bologna, Italy, 51-60.
12. Andrienko, G., Andrienko, N., and Savinov, A., Choropleth Maps: Classification revisited, In Proceedings ICA 2001, Beijing, China, Vol. 2, pp. 1209-1219.
13. Andrienko, N., Andrienko, G., Savinov, A., Voss, H., and Wettschereck, D., Exploratory Analysis of Spatial Data Using Interactive Maps and Data Mining, Cartography and Geographic Information Science 2001, v.28 (3), pp. 151-165