Mining geospatial data in a transductive setting

A. Appice, N. Barile, M. Ceci, D. Malerba & R. P. Singh
Dipartimento di Informatica, Università degli Studi di Bari, Italy

Abstract

Many organizations collect large amounts of spatially referenced data. Spatial Data Mining targets the discovery of interesting, implicit knowledge from such data. The specific classification task has been extensively investigated in the classical inductive setting, where only labeled examples are used to generate a classifier, discarding a large amount of information potentially conveyed by the unlabeled instances to be classified. In this work spatial classification is based on transduction, an inference mechanism "from particular to particular" which uses both labeled and unlabeled data to build a classifier whose main goal is that of classifying (only) unlabeled data as accurately as possible. The proposed method, named TRANSC, employs a principled probabilistic classification in multi-relational data mining to face the challenges posed by handling spatial data. The predictive accuracy of TRANSC has been evaluated on two real-world spatial datasets.

1 Introduction

The expanding market for spatial databases and Geographic Information System (GIS) technologies is driven by the pressure from the public sector, environmental agencies and industries to provide innovative solutions to data applications that involve spatial data, that is, a collection of (spatial) objects organized in thematic layers (e.g., roads, rivers). A thematic layer is characterized by a geometrical representation as well as several non-spatial attributes, called thematic attributes. A GIS provides the set of functionalities to store, retrieve and manage both geometrical representation and thematic attributes stored in a spatial database. Anyway, the range of GIS applications can be extended by adding spatial data mining facilities to the systems [1] to extract implicit knowledge from georeferenced data.

WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070141

Classification in Spatial Data Mining has been extensively investigated in the classical inductive setting, where the goal is to estimate the value of the unknown underling classification function g at a given set of points of a working sample W based on only the training sample S. The usual approach to estimating these class values consists in first finding an approximation g' to the desired function g and then using this approximation to get the required estimates. However, this approach is not always appropriate when the cardinality of the training sample S is much smaller than that of the working sample W; this is often the case in many real-world situations, such as several geographical data mining applications, where large amounts of unlabeled geographical objects (e.g., map cells) are available and manual annotation is therefore fairly expensive. The main limitation of inductive approaches is that only labeled examples are used to generate a classifier, discarding a large amount of information potentially conveyed by the unlabeled instances to be classified. The idea of *transductive inference* (or *transduction*) [2] is to analyze both the labeled (training) data S and the unlabeled (working) data W to build a classifier and classify (only) the unlabeled data W as accurately as possible.

In the literature, several transductive learning methods have been proposed for support vector machines [3, 4], k-NN classifiers [5] as well as general classifiers [6]. However, all of those transductive learning algorithms assume training/working data are represented as a single table (or database relation) whose rows (or tuples) represent independent units of the sample population, while columns correspond to properties of these units (*single table assumption*). This tabular representation of data, also known as *propositional* representation, turns out to be too restrictive for complex applications such as spatial applications where, different spatial objects may have distinctive properties, which are properly modeled by as many data tables as the number of object types. Moreover, the attributes of spatial objects in the neighborhood may affect each other (spatial autocorrelation), hence the need for representing object interactions by additional data tables.

In this paper, we propose a novel transductive classification algorithm, named TRANSC (TRANsductive Structural Classifier), which exploits the expressive power of Multi-Relational Data Mining (MRDM) [7] to deal with spatial data in their original form. Knowledge on data model (e.g., foreign keys) is obtained "free of charge" from the database schema used as a guide in the search process. TRANSC performs a probabilistic classification based on the multi-relational extension of the naïve Bayesian classifier.

2 Background and related work

Naïve Bayesian classifiers have been designed to perform probabilistic classification tasks. Given a feature-vector representation of a test data x, a classical naïve Bayesian classifier assigns x to the class C_i that maximizes the *posterior probability* $P(C_i|x)$. According to the Bayes theorem, $P(C_i|x)$ can be expressed as $P(C_i|x) = P(C_i)P(x|C_i)/P(x)$. Under the conditional

independence (or *naïve*) assumption of object attributes, the likelihood $P(x|C_i)$ can be factorized as $P(x|C_i) = P(x_1, \ldots, x_m|C_i) = P(x_i|C_i) \times \ldots \times P(x_m|C_i)$, where $x_1, \ldots, x_m$ represent the attribute values different from the class label used to describe the object x. Naïve Bayesian classifiers have been proved accurate even when the conditional independence assumption is grossly violated [8].

The above formalization of a naïve Bayesian classifier only applies to propositional representations. In the case of relational representations, some extensions of it are necessary. The basic idea is that of using a set of relational patterns to describe an object to be classified, and then to define a suitable decomposition of the likelihood $P(x|C_i)$ *à la* naïve Bayesian classifier to simplify the resolution of the probability estimation problem. Each $P(x|C_i)$ is computed on the basis of a set $\Re = \{AN_j \Rightarrow CO_j\}$ of relational classification rules, where the consequent CO_j represents the class label for an object X and the antecedent AN_j is a conjunction of literals describing both relations and properties of objects. More precisely, if $\Re(x) \subseteq \Re$ is the set of rules whose antecedent covers the reference object x, then $P(x|C_i) = P(\bigwedge_{R_k \in \Re(x)} antecedent(R_k)|C_i)$. This extension of the naïve Bayesian classifier is adopted in the (inductive) multi-relational naïve Bayesian classifier Mr-SBC [9] we base our approach on.

In *semi-supervised learning*, labeled and unlabeled data are used for training but the inferential principle is still inductive, that is, general rules hopefully valid for the whole instance space are generated. A semi-supervised learning algorithm was proposed in [10], where the naïve Bayesian classifier is combined with the EM algorithm. Vapnik [2] was the first to introduce the idea of transductive learning with his transductive Support Vector Machines (SVMs). Blum and Chawla [11] proposed to use a similarity measure to construct a graph and then partitions the graph in such a way that it (roughly) minimizes the number of similar pairs of examples that are given different labels. An evolution of this work is the transductive version of k-NN, which was designed to avoid the myopia of the greedy search strategy adopted in graph partitioning by solving an optimization problem via spectral methods [5].

3 Probabilistic transduction in TRANSC

The problem of transductive classification solved by TRANSC can be formulated as follows:
Given:
• a database schema S which consists of a set of h relational tables $\{T_0, \ldots, T_{h-1}\}$, a set PK of primary key constraints on the tables in S, and a set FK of foreign key constraints on the tables in S
• a target relation $T \in S$ and a target discrete attribute y in T, different from the primary key of T, whose domain is the finite set $\{C_1, C_2, \ldots, C_L\}$
• the projection T' of T on all attributes of T except y
• a training (working) set that is an instance TS (WS) of the database schema S with known values for y

Find: the most accurate prediction of y for examples in WS represented as a tuple of $t \in WS.T'$ and all tuples related to t in WS according to FK.

In keeping with the main idea adopted in [4], we iteratively try to change the classification of "borderline" cases, that is, training and working examples that are more likely subject to classification errors. In particular, we propose an algorithm (see Algorithm 1) which starts with a given classification and, at each iteration, alternates a step during which examples are reclassified and a step during which the class of "borderline" examples is changed.

Algorithm 1 Top level transductive algorithm description

```
1: transductiveClassifier(initialClassification, TS, WS)
2: classification1 ← initialClassification; changedExamples ← φ; i ← 0;
3: repeat
4:    prevClassification ← classification1;
5:    prevChangedExamples ← changedExamples;
6:    classification2 ← reclassifyExamplesKNN(classification1, TS, WS);
7:    (classification1, changedExamples) ← changeClass(classification2);
8: until ( (computeOverlap(prevChangedExamples,changedExamples) ≥ MAX-
   OVERLAP) OR (++i ≥ MAX_ITERS) )
9: return prevClassification
```

The initial classification of $E \in WS \cup TS$ is obtained according to the function $preclass(E)$ that returns $class(E)$ if $E \in TS$, $BayesianClassification(E)$, otherwise. $BayesianClassification(E)$ is the initial inductive classifier built from the training set TS. The initial classifier is obtained by means of an improved version of the relational probabilistic learning algorithm Mr-SBC [9] whose search strategy is enhanced by considering cyclic paths in the set of foreign keys FK and whose discretization is now performed by means of an equal-width based strategy.

The examples are then reclassified by means of a variant of the k-NN algorithm tailored for transductive inference in MRDM. The idea is to classify each example $E \in WS \cup TS$ on the basis of a k-sized neighborhood $N_k(E) = \{E_1, \ldots, E_k\}$ consisting of the k examples of $WS \cup TS$ closest to E with respect to a dissimilarity measure d. This step aims at identifying the value y' of the L-dimensional class probability vector associated to E, that is $y' = (y_1(E), \ldots, y_L(E))$, where each $y_i(E) = P(class(E) = C_i)$ is estimated based on $N_k(E)$.

Each probability $P(class(E) = C_i)$, $i = 1, \ldots, L$ is estimated as follows:

$$P(class(E) = C_i) = |\{E_j \in N_k(E) | C_{E_j} = C_i\}|/k \tag{1}$$

where C_{E_j} is the class value associated to E_j at the previous step; at the first step, C_{E_j} is the class label returned by $preclass(E_j)$. It should be noted that $P(class(E) = C_i)$ is estimated according to the transductive inference principle, as both the training and the working sets are taken into account in the process.

The $changeClass$ procedure is in charge of changing the classification of borderline examples. Unlike what proposed in [4], where support vectors are used to identify examples on the border, we consider examples for which the entropy of the decision made by the classifier is maximum. The entropy for each example is computed from the probabilities associated with each class C_i:

$$Entropy(E) = -\sum_{i=1,\ldots,L} P(class(E) = C_i) \times log(P(class(E) = C_i)) \quad (2)$$

The examples are then ordered according to the entropy function and the class label of at most the first k examples having $Entropy(E)$ >*MINENTROPY* is changed. The class to which each selected example E is assigned is the most likely class C_i for E among those remaining after the old class of E has been excluded. The threshold k is necessary in order to avoid changing the class of several examples that would lead to erroneously change class of entire "clusters".

Two distinct stopping criteria are used. The first criterion stops the execution of the algorithm when the maximum number of iterations (*MAX_ITERS*) is reached. This guarantees the termination of the algorithm. Indeed, our experiments showed that this criterion is rarely attained when *MAX_ITERS* is as small as 10. The second criterion stops the execution when the examples processed in an iteration remain the same. For this purpose, the overlap between two sets of examples is determined. The $computeOverlap$ function returns the ratio between the cardinality of the intersection between the sets of examples and that of their union.

The classifier returned by Mr-SBC starting from the training set TS is not just employed to pre-classify the working examples in WS. Indeed, the initial Mr-SBC classifier includes a set of first-order classification rules used to represent the examples to be classified. TRANSC reuses such rules to derive a boolean feature-vector representation of each example in WS on which the similarity function subsequently determined is based. More formally, let $\Re$ be the set of classification rules extracted by Mr-SBC in the form: $p_0(A_1, y) \leftarrow p_1(A_1, A_2)$, $p_2(A_2, A_3)$, $\ldots$, $p_{s-1}(A_{s-1}, A_s)$, $p_s(A_s, c)$ where:

• s is at most *MAX_LEN_PATH*, a user defined parameter which limits the number of Mr-SBC refinement steps.

• p_0 is a predicate associated to the target table T and to the target attribute y.

• $p_l, \quad l = 1, \ldots, s-1$, is a predicate associated to a table $T_{i_l} \in S$ such that a foreign key exists in S between T_{i_l} and $T_{i_{l-1}}$.

• p_s is an *optional* property predicate associated to $T_{i_{s-1}}$ and to one of its attributes.

Rules in $\Re$ are not used by TRANSC in the form returned by Mr-SBC. Indeed, the discretization of continuous attributes performed by Mr-SBC leads to generate rules for which the p_s predicate is in the form $T_{i_{s-1}}.Attr \in [v_1, v_2]$, where $[v_1, v_2]$ is a bin. However, this representation may cause information loss on the order relation of continuous values. To overcome this problem, we follow the idea formulated in [12] and we transform p_s in $T_{i_{s-1}}.Attr \leq v_2$. The new set of rules $\Re'$ obtained in this way permits to consider more similar two examples whose feature values appear in two consecutive bins than two examples whose feature values are

far apart. Once $\Re'$ has been constructed, each example $E \in WS$ is described by a boolean feature-vector V_E composed by $|\Re'|$ features (one for each rule). If the antecedent of a rule $(AN_j \Rightarrow CO_j) \in \Re'$ *covers* E, that is, a substitution θ exists such that $AN_j\theta \subseteq E$, then the j-th element of V_E is set to $true$; $false$ otherwise.

The similarity between two examples E_1 and E_2 is determined by means of the Kendall, Sokal-Michener (1958) similarity measure [12]:

$$s(E_1, E_2) = cardinality(V_{E_1} \quad XNOR \quad V_{E_2})/|\Re'| \tag{3}$$

where $cardinality(\bullet)$ returns the number of $true$ values included in a boolean vector. Coefficient (3) takes values in the unit interval: $s(E_1, E_2) = 1$ if the two vectors match perfectly, while $s(E_1, E_2) = 0$ if the two vectors are orthogonal. The dissimilarity between two examples is: $d(E_1, E_2) = 1 - s(E_1, E_2)$.

4 Experiments

We compared TRANSC and Mr-SBC to empirically validate the transductive reformulation of the relational probabilistic classification on real-world spatial applications when few labeled examples are available.

The two algorithms are compared on the basis of the average misclassification error on the same K-fold cross validation of each dataset. For each trial (fold), both TRANSC and Mr-SBC are trained on a single database and tested on the hold-out $K - 1$ database instances forming the working set. The error rates reported in this work are significantly higher than those reported in [9] [13] due to the peculiar experimental design (a small training set and a large working set).

Since the performance of TRANSC may vary significantly depending on the size (k) of the neighborhood, experiments for different k are performed in order to set the optimal value. In theory, we should experiment with each value of k. However, as observed in [14], the search for the optimal k can be substantially reduced from $[1, |D|]$ to $[1, \sqrt{|D|}]$, without loosing too much accuracy. Hence, we have decided to consider only $k = \eta i$ such that i ranges in $[1, \sqrt{|D|}/h]$ and η is a step value. The classifiers mined in this study are obtained by setting *MAX_LENGTH_PATH=3*, *MAX_ITERS=10*, *MINENTROPY*=0.65 and *MAXOVERLAP*=0.5. The step value η differs for each dataset.

We have tested our transductive algorithm on two different spatial data collections, that is, the North-West England Census Data and the Munich Census Data.

The North-West England Census data are obtained from both census and digital maps data provided by the European project SPIN! (http://www.ais.fraunhofer.de/KD/SPIN/project.html). Data concern Greater Manchester, one of the five counties of North West England (NWE). Greater Manchester is divided into 214 census sections (wards). Census data are available at ward level and provide socio-economic statistics (e.g. mortality rate) as well as some measures of the deprivation of each ward according to information provided by Census combined into single index scores. The goal of the classification task is to predict

the value of the Jarman index (low or high value) deprivation factor by exploiting both other deprivation factors (Townsend index, Carstairs index and DoE index), mortality rate and geographical factors represented in topographic maps of the area. Vectorized boundaries of the 1998 census wards as well as of other Ordnance Survey digital maps of NWE are available for several layers such as urban area (115 lines), green area (9 lines), road net (1687 lines), rail net (805 lines) and water net (716 lines). The objects on each layer have been stored as tuples of relational tables including information on the object type (TYPE). For instance, an urban area may be either a "large urban area" or a "small urban area". Topological relationships between wards and objects in all these layers are materialized as relational tables expressing non-disjoint relations. The number of materialized "non disjoint" relationships is 5313.

The Munich Census Data concern the level of monthly rent per square meter for flats in Munich expressed in German Marks (http://www.di.uniba.it/∼ceci/micFiles/munich_db.tar.gz). The data describe 2180 geo-referenced flats located in the 446 subquarters of Munich obtained by first dividing the Munich metropolitan area up into three areal zones and then by decomposing each of these zones into 64 districts. The vectorized boundaries of subquarters, districts and zones as well as the map of public transport stops consisting of public train stops (56 subway (U-Bahn) stops, 15 rapid train (S-Bahn) stops and 1 railway station) within Munich are available for this study. The objects included in these layers are stored in different relational tables (SUBQUARTERS, TRANSPORT_STOPS and FLATS). Information on the "area" of subquarters is stored in the corresponding table. Transport stops are described by means of their type (U-Bahn, S-Bahn or Railway station), while flats are described by means of their "monthly rent per square meter", "floor space in square meters" and "year of construction". The target attribute was represented by the "monthly rent per square meter", whose values have been discretized into the two values $low = [2.0, 14.0]$ or $high =]14.0, 35.0]$. The spatial arrangement of data is defined by both the "close_to" relation between Munich metropolitan subquarters areas and the "inside" relation between public train stops and metropolitan subquarters. Both of these topological relations are materialized into relational tables (CLOSE_TO and INSIDE).

The average misclassification errors of TRANSC and Mr-SBC are reported in Tables 1 and 2. The results are obtained according to both a 10-fold cross validation (CV) of the data and a 20-fold CV of the same data. In the case of the NWE Census Data, we set $k \in \{4, 7, 9, 11, 14\}$, while in the case of the Munich Census Data we set $k \in \{9, 18, 27, 36, 45\}$. In both datasets, results confirm an improved accuracy for the transductive setting with respect to the inductive one. The gain depends on the k value and this result is more evident in the case of 10-fold CV. In 20-fold CV, there is an error propagation through algorithm iterations due to the presence of few training examples. A deeper analysis of results of 10-fold CV reveals that accuracy increases with high values of k (k=11 for NWE and k= 36 for Munich), but at the some time accuracy decreases when k approximates $\sqrt{|D|}$. This poses the problem of determining some criterion to automatically approximate best k value. Finally, the accuracy is also affected by the number of bins processed by

Table 1: TRANSC vs. Mr-SBC on NWE Census Data: average misclassification error on the working sets. Number of bins (Nb) in Mr-SBC discretization is set to 10.

Experiment	TRANSC					Mr-SBC
	k=4	k=7	k=9	k=11	k=14	
Avg 10-CV Error	23.38%	21.10%	19.79%	18.04%	19.08%	22.71%
%error loss	-2.97%	7.06%	12.84%	20.56%	15.99%	
Avg 20-CV Error	33.87%	34.41%	33.82%	33.20%	33.28%	34.31%
%error loss	0.00%	-1.60%	0.15%	1.96%	1.75%	

Table 2: TRANSC vs. Mr-SBC on Munich Census Data: average misclassification error on the working sets. Nb=40.

Experiment	TRANSC					Mr-SBC
	k=9	k=18	k=27	k=36	k=45	
Avg 10-CV Error	28.99%	28.61%	28.36%	28.30%	28.15%	31.23%
%error loss	7.17%	8.41%	9.19%	9.40%	9.86%	
Avg 20-CV Error	37.25%	36.30%	36.73%	36.67%	36.78%	37.79%
%error loss	1.44%	3.94%	2.81%	2.98%	2.68%	

Mr-SBC. Some results, which we omit here due to space limitations, empirically prove that the higher the gain the lower the number of bins.

5 Conclusions

In this work we have investigated the combination of transductive inference with principled probabilistic MRDM classification for tasks of spatial classification. Our proposal consists in inducing an initial multi-relational naïve Bayesian classifier (Mr-SBC) by processing only training (i.e., labeled) examples and then using this classifier to preliminarily label the working (i.e., unlabeled) data. The initial classification of the working examples is then refined iteratively over a finite number of steps, each of which consists in a k-NN classification of unlabeled (working) examples and a subsequent reclassification of some "borderline" examples. Neighbors are determined by computing a distance measure on a propositionalized representation of working examples. Propositional features are obtained by transforming multi-relational rules mined with Mr-SBC in boolean

features. This transductive classifier (TRANSC) has been compared to its inductive counterpart (Mr-SBC) on two spatial datasets. Experiments are in favor of TRANSC and the percentage of accuracy improvement of the transductive setting with respect to the inductive one appears better than the improvement observed in [5] when comparing SVMs in both the inductive and transductive propositional setting.

Acknowledgement

This work partially fulfills the research objective of ATENEO-2006 project entitled "Metodi di scoperta di conoscenza per ubiquitous computing".

References

[1] Koperski, K., *Progressive Refinement Approach to Spatial Data Mining*. Ph.D. thesis, Computing Science, Simon Fraser University, Canada, 1999.
[2] Vapnik, V., *Statistical Learning Theory*. Wiley: New York, 1998.
[3] Gammerman, A., Azoury, K. & Vapnik, V., Learning by transduction. *UAI 1998*, Morgan Kaufmann, pp. 148–155, 1998.
[4] Joachims, T., Transductive inference for text classification using support vector machines. *ICML 1999*, Morgan Kaufmann, pp. 200–209, 1999.
[5] Joachims, T., Transductive learning via spectral graph partitioning. *ICML 2003*, Morgan Kaufmann, 2003.
[6] Kukar, M. & Kononenko, I., Reliable classifications with machine learning. *ECML 2002*, Springer-V., pp. 219–231, 2002.
[7] Džeroski, S. & Lavrač, N., *Relational Data Mining*. Springer-V., 2001.
[8] Domingos, P. & Pazzani, M., On the optimality of the simple bayesian classifier under zeo-ones loss. *Machine Learning*, **28(2-3)**, pp. 103–130, 1997.
[9] Ceci, M., Appice, A. & Malerba, D., Mr-SBC: a multi-relational naive bayes classifier. *PKDD 2003*, Springer-V., volume 2838 of *LNAI*, pp. 95–106, 2003.
[10] Nigam, K., McCallum, A.K., Thrun, S. & Mitchell, T.M., Text classification from labeled and unlabeled documents using EM. *Machine Learning*, **39(2/3)**, pp. 103–134, 2000.
[11] Blum, A. & Chawla, S., Learning from labeled and unlabeled data using graph mincuts. *ICML 2001*, Morgan Kaufmann, pp. 19–26, 2001.
[12] Esposito, F., Malerba, D., Tamma, V. & Bock, H., *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag New York, Inc., 2000.
[13] Ceci, M. & Appice, A., Spatial associative classification: propositional vs structural approach. *Journal of Intelligent Information Systems*, **27(3)**, pp. 191–213, 2006.
[14] Gora, G. & Wojna, A., RIONA: A classifier combining rule induction and k-nn method with automated selection of optimal neighbourhood. *ECML 2002*, Springer-V., volume 2430 of *LNAI*, pp. 111–123, 2002.