Transductive Learning from Relational Data

Michelangelo Ceci, Annalisa Appice, Nicola Barile, and Donato Malerba

Dipartimento di Informatica, Università degli Studi di Bari
via Orabona, 4 - 70126 Bari - Italy
{ceci,appice,malerba}@di.uniba.it, n.barile@gmail.com

Abstract. Transduction is an inference mechanism "from particular to particular". Its application to classification tasks implies the use of both labeled (training) data and unlabeled (working) data to build a classifier whose main goal is that of classifying (only) unlabeled data as accurately as possible. Unlike the classical inductive setting, no general rule valid for all possible instances is generated. Transductive learning is most suited for those applications where the examples for which a prediction is needed are already known when training the classifier. Several approaches have been proposed in the literature on building transductive classifiers from data stored in a single table of a relational database. Nonetheless, no attention has been paid to the application of the transduction principle in a (multi-)relational setting, where data are stored in multiple tables of a relational database. In this paper we propose a new transductive classifier, named TRANSC, which is based on a probabilistic approach to making transductive inferences from relational data. This new method works in a transductive setting and employs a principled probabilistic classification in multi-relational data mining to face the challenges posed by some spatial data mining problems. Probabilistic inference allows us to compute the class probability and return, in addition to result of transductive classification, the confidence in the classification. The predictive accuracy of TRANSC has been compared to that of its inductive counterpart in an empirical study involving both a benchmark relational dataset and two spatial datasets. The results obtained are generally in favor of TRANSC, although improvements are small by a narrow margin.

1 Introduction

In the usual inductive classification setting, data is supposed to have been generated independently and identically distributed (i.i.d.) from an unknown probability distribution P on some domain X and are labeled according to an unknown function g. The domain of g is spanned by m independent (predictor) random variables X_i (either numerical or categorical), that is, $X = X_1, X_2, \ldots, X_m$. The range of g is a finite set $Y = \{C_1, C_2, \ldots, C_L\}$, where each C_i is a distinct class label. After being inputted a training sample $S = \{(x, y) \in X \times Y | y = g(x)\}$, an inductive learning algorithm returns a function f that is hopefully close to g on the domain X. However, there are many cases in which the goal is to estimate the value of the unknown function g at a given set of points of a working

P. Perner (Ed.): MLDM 2007, LNAI 4571, pp. 324–338, 2007.

sample $W \subseteq X$ based on the training sample S. The usual way of estimating these values consists in first finding an approximation g' to the desired function g and then using this approximation to get the required estimates. This approach is not always the best when the cardinality of the training sample S is much smaller than that of the working sample W, which is often the case in many real-world situations. It characterizes the traditional inductive learning setting, which uses only labeled examples to generate a classifier and discards a large amount of information potentially conveyed by the unlabeled instances to be classified. Conversely, the idea of *transductive inference* (or *transduction*) [20] is to analyze both the labeled (training) data S and the unlabeled (working) data W to build a classifier whose main goal is that of classifying (only) the unlabeled data W as accurately as possible.

Several transductive learning methods have been proposed in the literature for support vector machines [1] [10] [13] [6], for k-NN classifiers [14] and even for general classifiers [15]. However, despite the growing interest of the scientific community for transductive inference, all of those transductive learning algorithms are based on the *single-table assumption* [22], according to which the training/test data are represented in 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. This classic tabular representation of data, also known as *propositional* or *feature-vector* representation, turns out to be too restrictive for some complex applications. For instance, in spatial data mining, different spatial objects may have distinctive properties, which can be properly modeled by as many data tables as the number of object types. Moreover, attributes of the neighbors of spatial objects may affect each other (spatial autocorrelation), hence the need for representing object interactions by additional data tables. Although several methods have been proposed to transform a *(multi-)relational* (or *structural*) representation of training data into a single table, this approach (known as *propositionalization*) is fraught with many difficulties in practice [7,11].

In this paper, we propose a novel transductive classification algorithm, named TRANSC (TRANsductive Structural Classifier), that exploits the expressive power of Multi-Relational Data Mining (MRDM) to deal with relational data in their original form. This means that knowledge on the relational data model (e.g., foreign key constraints) is obtained free of charge from the database schema and used to guide the search process. The method works in a transductive setting and employs a probabilistic approach to classification. Information on the potential uncertainty of classification conveyed by probabilistic inference is useful when small changes in the attribute values of a test case may result in sudden changes of the classification. It is also useful when missing (or imprecise) information may prevent a new object from being classified at all [5].

The rest of the paper is organized as follows. In the next section, the background of this research and some related works are introduced, while the (multi-)relational transductive learning problem solved by TRANSC is formally defined in Section 3. In Section 4 experimental results are reported for both

a benchmark dataset typically used in MRDM and for two spatial datasets. Finally, Section 5 concludes and discusses ideas for further work.

2 Background and Related Work

The combination of relational representation with principled probabilistic and statistical approaches to inference and learning has been deeply investigated. In particular, relational 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* $P(C_i|x)$. By applying the Bayes theorem, $P(C_i|x)$ is expressed as follows:

$$P(C_i|x) = \frac{P(C_i)P(x|C_i)}{P(x)}. \tag{1}$$

Under the conditional independence (or *naïve*) assumption of object attributes, the likelihood $P(x|C_i)$ can be factorized as follows:

$$P(x|C_i) = P(x_1, \ldots, x_m|C_i) = P(x_i|C_i) \times \ldots \times P(x_m|C_i) \tag{2}$$

where $x_1, \ldots, x_m$ represent the attribute values different from the class label used to describe the object x. Surprisingly, naïve Bayesian classifiers have been proved accurate even when the conditional independence assumption is grossly violated. This is due to the fact that when the assumption is violated, although the estimates of posterior probabilities may be poor, the correct class still has the highest estimate. This leads to correct classifications [8].

The above formalization of a naïve Bayesian classifier is clearly limited to propositional representations. In the case of relational representations, some extensions 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 Bayes to simplify the resolution of the probability estimation problem.

An example of relational pattern considered in this work is the following:

$$molecule\_Atom(A, B) \wedge molecule\_Type(B, [22, 27]) \Rightarrow molecule\_Attribute(A, active).$$

This is a relational classification rule generated for the Mutagenesis dataset considered in Section 4.1. The literal $molecule\_Attribute(A, active)$ in the consequent of the rule represents the class label (i.e. "active") associated to the molecule A. The literal $molecule\_Atom(A, B)$ in the antecedent of the rule is a *structural characteristic* representing the foreign-key constraint between the tables *Molecule* and *Atom*, while the literal $molecule\_Type(B, [22, 27])$ is a *property* stating that the value of the attribute *Type* of the atom B (composing the molecule A) is a number in the interval [22,27].

Each $P(x|C_i)$ is computed on the basis of a set $\Re = \{A_j \Rightarrow y(X, C_i)\}$ of relational classification rules, where $C_i \in Y$, $y(\_, \_)$ is a binary predicate

representing the class label for an example X and the antecedent A_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:

$$P(x|C_i) = P(\bigwedge_{R_k \in \Re(x)} antecendent(R_k)|C_i). \quad (3)$$

This extension of the naïve Bayesian classifier to the case of multi-relational data was originally proposed by Pompe and Kononenko [18] and was recently reworked by Flach and Lachiche [9]. In both works, the conditional independence assumption is straightforwardly applied to all literals in $\bigwedge_{R_k \in \Re(x)} antecendent(R_k)$. However, this may lead to underestimate $P(x|C_i)$ when several similar rules in $\Re$ are considered for the class C_i. Therefore, in this study, we employ a less biased procedure for the computation of the probabilities 3, namely that adopted in the multi-relational naïve Bayesian classifier Mr-SBC [5].

All above mentioned works on relational naïve Bayesian classifiers ignore unlabeled data when mining the classifier. In *semi-supervised learning* approaches, both labeled and unlabeled data are used for training, but the inferential principle is still inductive, that is, a general rule hopefully valid for the whole instance space is generated. An example of semi-supervised learning algorithm has been proposed by Nigam et al. [16], who combine the the naïve Bayesian classifier with the Expectation-Maximization (EM) algorithm. The former is trained on labeled data and provides an initial classification of unlabeled data, while the latter is used to perform hill-climbing in data likelihood space, finding the classifier parameters that locally maximize the likelihood of all the data, both the labeled and the unlabeled.

Vapnik [20] has introduced the transductive Support Vector Machines (SVMs), which take into account a particular test set and try to minimize the misclassification rate of just those particular examples. A different approach has been proposed by Blum and Chawla [2], who uses a similarity measure to construct a graph and then partitions the graph in such a way that it minimizes (roughly) 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 has been designed to avoid the myopia of the greedy search strategy adopted in graph partitioning by efficiently and globally solving an optimization problem via spectral methods [14].

Finally, some studies on transductive inference have investigated the opportunity of applying transduction to evaluate the predictive reliability of a real-valued regression model. The basic idea in [3] is to construct transductive predictors and to establish a connection between initial and transductive predictions. An initial predictor is obtained as the model that best fits the training set. It is used to assign a label to a single unlabeled example to be included in the training set and the new training set is used to obtain the final transductive predictor in an iterative process.

3 Probabilistic Transduction in TRANSC

Let $D = \{(x, y) \in X \times Y | y = g(x)\}$ be a dataset labeled according to an unknown function g whose range is a finite set $Y = \{C_1, C_2, \ldots, C_L\}$. Our transductive classification problem is formalized as follows:

Input
• a training set $S \subset D$ and
• the projection of the working set $W = D - S$ on X;
Output: a prediction of the class value (y) of each example in the working set W which is as accurate as possible.

The learner receives full information (including labels) on the examples in S and partial information (only that concerning the independent variables X_i) on the examples in W and is required to predict the class values only of the examples that W consists of. The original formulation of the problem of function estimation in a transductive (*distribution-free*) setting requires that S be sampled from D without replacement. This means that, unlike the standard inductive setting, the examples in the training (and working) set are supposed to be mutually dependent. Vapnik also introduced a second (*distributional*) transduction setting in which the learner receives training and working sets, which are assumed to be drawn i.i.d. from some unknown distribution. As shown in [20] (Theorem 8.1), error bounds for learning algorithms in the distribution-free setting also apply to the more popular distributional transductive setting. Therefore, in this work we focus our attention to the first setting.

In the case of relational data, the problem of transductive classification we aim at solving 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 $T \in S$
- 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 the values 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.

This problem is solved by TRANSC by accessing, as in the propositional case, both the full representation of instances in the training set (including that of y) and the partial representation of instances in the working set (represented by T' and its joined tables).

In keeping with the main idea adopted in [13], we iteratively refine the classification by changing the classification of training and working examples in the

"borderline" of the class that would be more likely subject to 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;
3: changedExamples ← φ;
4: i ← 0;
5: repeat
6:    prevClassification ← classification1;
7:    prevChangedExamples ← changedExamples;
8:    classification2 ← reclassifyExamplesKNN(classification1, TS, WS);
9:    (classification1, changedExamples) ← changeClassification(classification2);
10: until ( (++i ≥ MAX_ITERS) OR
    (computeOverlap(prevChangedExamples,changedExamples) ≥ MAXOVERLAP))
11: return prevClassification
```

The initial classification of an example $E \in WS \cup TS$ is obtained according to the following classification function:

$$preclass(E) = \begin{cases} class(E) & \text{if } E \in TS \\ BayesianClassification(E) & \text{if } E \in WS \end{cases}$$

where $BayesianClassification(E)$ is the classification function corresponding to the initial inductive classifier built from the training set TS. Such an initial classifier is obtained by means of an improved version of the relational probabilistic learning algorithm Mr-SBC [5] whose search strategy is enhanced by considering cyclic paths in the set of foreign keys FK.

The examples are then reclassified by means of a version 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 included in $WS \cup TS$ 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 the example 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 $P(class(E) = C_i)$ is estimated as follows:

$$P(class(E) = C_i) = \frac{|\{E_j \in N_k(E) | C_{E_j} = C_i\}|}{k} \tag{4}$$

such that:

- $P(class(E) = C_i) \geq 0$ for each $i = 1, \ldots, L$,
- $\sum_{i=1,\ldots,L} P(class(E) = C_i) = 1$.

In Equation (4), C_{E_j} is the generic class value associated to the example 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 training and working set are taken into account in the process.

The $changeClassification$ procedure is in charge of changing the classification of the examples on the borderline of a class. Unlike what proposed in [13], where support vectors are used to identify examples on the border, in our case we consider examples for which the entropy of the decision taken by the classifier is maximum. The entropy for each example E is computed from the probabilities associated with each class C_i:

$$Entropy(E) = -\sum_{i=1,\ldots,L} P(class(E) = C_i) \times log(P(class(E) = C_i)) \quad (5)$$

The examples are 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 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".

In Algorithm 1, 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 the parameter MAX_ITERS is as small as 10.

The second criterion aims at stopping execution when a cycle insists on the same examples of the previous one. 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 the cardinality 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 $\Re = \{A_j \Rightarrow y(X, C_i)\}$ be the set of classification rules extracted by Mr-SBC, where $C_i \in Y$, $y(\_, \_)$ is a binary predicate representing the class label for an example X and the antecedent A_j is the conjunction of at most MAX_LEN_PATH literals describing both relations and properties of objects. Then each example $E \in WS$ is described by a boolean feature-vector

V_E composed by $|\Re|$ elements, that is, $A_1, \ldots, A_{|\Re|}$. If the antecedent of a rule $(A_j \Rightarrow y(X, C_i)) \in \Re$ *covers* E, that is, a substitution θ exists such that $A_j\theta \subseteq E$, then the j-th element of V_E is set to *true*; otherwise, it is set to $false$.

The similarity between two examples E_1 and E_2 is determined by matching the *true* values of the corresponding vectors V_{E_1} and V_{E_2}. More precisely, by computing Jaccard's similarity coefficient, which is defined as follows:

$$s(E_1, E_2) = \frac{cardinality(V_{E_1}\ AND\ V_{E_2})}{cardinality(V_{E_1}\ OR\ V_{E_2})} \tag{6}$$

where $cardinality(\bullet)$ returns the number of *true* values included in a boolean vector. Coefficient 6 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 or in the degenerate case of no *true* value occurring in both vectors. The dissimilarity between two examples is then defined as follows:

$$d(E_1, E_2) = 1 - s(E_1, E_2) \tag{7}$$

4 Experiments

An empirical evaluation of our algorithm was carried out on both the Mutagenesis dataset, which have been used extensively in testing MRDM algorithms, and on two real-world spatial data collections concerning North West England Census data and Munich Census data, respectively.

We compared the performance of TRANSC to that of Mr-SBC in order to identify the advantages of employing a transductive reformulation of the problem of relational probabilistic classification in real-world applications where few labeled examples are available and manual annotation is fairly expensive.

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 dataset, the target table is first divided into K blocks of nearly-equal size and then a subset of tuples related to the tuples of the target table block by means of foreign key constraints are extracted. This way, K database instances are created. For each trial, 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. It should be noted that the error rates reported in this work are significantly higher than those reported in other literature [5] [4] because of this peculiar experimental design. Indeed, unlike the standard cross-validation approach, here one fold at a time is set aside to be used as the *training set* (and not as the *test set*). Small training set sizes allows us to validate the transductive approach but result in high error rates as well.

A non-parametric Wilcoxon two-sample paired signed rank test [17] is employed to perform a pairwise comparison of the two algorithms. In this test, the summations on both positive (W+) and negative (W-) ranks determine the winner.

It should be noted that in our experiments the size of the working set is one order of magnitude greater than the size of the training set; this is something

rather different from what usually happens when testing algorithms developed according to the inductive paradigm. Since the performance of the transductive classifier TRANSC may vary significantly depending on the size (k) of the neighborhood used to predict the class value of each working example, experiments for different k are performed in order to set the optimal value. In theory, we should experiment with each value of k ranging in the interval $[1, |D|]$ where D is the labeled data set. However, as observed in [21] it is not necessary to consider all possible values of k during cross-validation to obtain the best performance. The best performances are obtained by means of cross-validation on no more than approximately ten values of k. A similar consideration has also been reported in [12], where it is shown that the search for the optimal k can be substantially reduced from $[1, |D|]$ to $[1, \sqrt{|D|}]$, without loosing too much accuracy of the approximation. Hence, we have decided to consider in our experiments only $k = \eta i$ such that i value ranges on the sample $[1, \sqrt{|D|}/h]$ and η is the step value.

Classifiers mined in all experiments in this study are obtained by setting $MAX\_LENGTH\_PATH = 3$, $MAX\_ITERS = 10$, $MINENTROPY = 0.65$ and $MAXOVERLAP = 0.5$. The step η is different for each dataset.

4.1 Benchmark Relational Data Application

The Mutagenesis dataset concerns the problem of identifying some mutagenic compounds. We have considered, similarly to most experiments on data mining algorithms reported in literature, the "regression friendly" dataset consisting of 188 molecules. A study on this dataset [19] has identified five levels of background knowledge. Each subset is constructed by augmenting a previous subset and provides richer descriptions of the examples. Table 1 shows the first three sets of background knowledge, the ones we have used in our experiments, where $BK_i \subset BK_{i+1}$ for $i = 0, 1$. The larger the background knowledge set, the more complex the learning problem. All experiments consist in a 10-fold cross validation ($K = 10$).

Table 1. Background knowledge for Mutagenesis data

Background	Description
BK_0	Data obtained with the molecular modeling package QUANTA. For each compound it obtains the atoms, bonds, bond types, atom types, and partial charges on atoms.
BK_1	Definitions in BK_0 plus indicators *ind1* and *inda* in molecule table.
BK_2	Variables (attributes) *logp* and *lumo* are added to definitions in BK_1.

The predictive accuracy of TRANSC was measured by considering the values $k \in \{2, 4, 6, 8, 10, 12\}$. For each setting BK_i ($i = 0, 1, 2$), the average misclassification error of both TRANSC and Mr-SBC is reported in Figure 1. Results show that with BK_0, TRANSC performs better than Mr-SBC, although the improvement is not statistically significant (see Table 2). The results in the BK_1 and

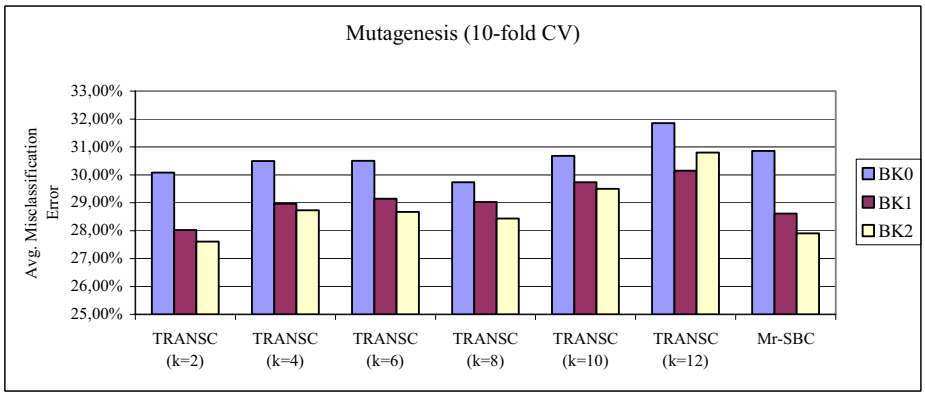

Fig. 1. TRANSC vs. Mr-SBC: average misclassification error on the working sets of Mutagenesis 10-CV data

BK_2 settings suggest different conclusions. As also shown in [5], the predictive accuracy of Mr-SBC increases so significantly when background knowledge is increased (BK_1 and BK_2 setting), that the consideration of unlabeled examples in a neighborhood can even lead to a deterioration in predictive accuracy. In this case, we obtain the best results when k is the lowest.

Table 2. Mutagenesis dataset: results of the Wilcoxon test (p-value) on average accuracy of TRANSC vs. Mr-SBC. The statistically significant p-values (< 0.05) are in italics. The sign + (-) indicates that TRANSC outperforms Mr-SBC (or vice-versa).

BK/k	2	4	6	8	10	12
BK_0	0.23 (+)	0.65 (+)	0.73 (+)	0.19 (+)	0.84 (+)	0.25 (-)
BK_1	0.42 (+)	0.65 (-)	0.76 (-)	0.55 (-)	0.35 (-)	0.2 (-)
BK_2	1.0 (+)	0.13 (-)	0.38 (-)	0.64 (-)	*0.02* (-)	*0.001* (-)

4.2 Spatial Data Application

We have also 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). These data concern Greater Manchester, one of the five counties of North West England (NWE). Greater Manchester is divided into ten metropolitan districts, each of which is in turn decomposed into censual sections (wards), for a total of two hundreds and fourteen wards. Census data are available at ward level and provide socio-economic statistics

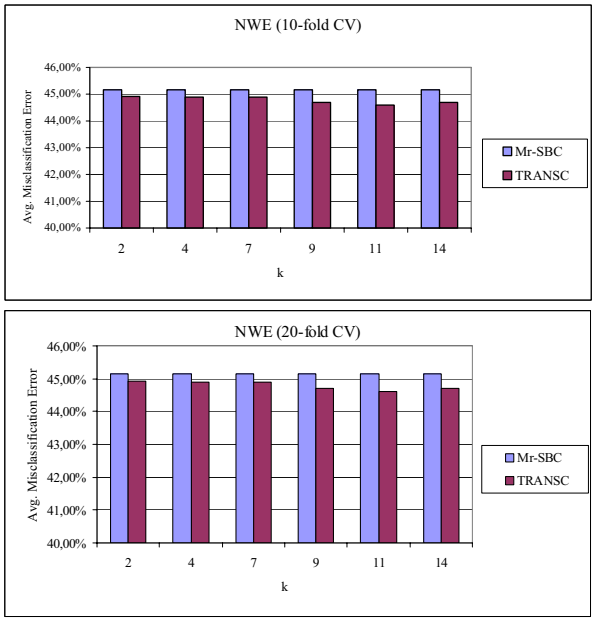

Fig. 2. TRANSC vs. Mr-SBC on NWE census data: average misclassification error on the working sets for 10-fold and 20-fold cross-validation

(e.g. mortality rate – the percentage rate of deaths with respect to the number of inhabitants) as well as some measures of the deprivation of each ward according to information provided by Census combined into single index scores. We have employed Jarman Underprivileged Area Score (which is designed to estimate the need for primary care), the indices developed by Townsend and Carstairs (used to perform health-related analyses), and the Department of the Environment's (DoE) index (which is used in targeting urban regeneration funds). The higher the index value the more deprived the ward. The mortality percentage rate takes values in the finite set $\{low = [0.001, 0.01], high =]0.01, 0, 18]\}$.

The goal of the classification task is to predict the value of the mortality rate by exploiting both deprivation factors and geographical factors represented in some linked topographic maps. Spatial analysis is possible thanks to the availability of vectorized boundaries of the 1998 census wards as well as of other Ordnance Survey digital maps of NWE, where several interesting layers such as urban area (115 lines), green area (9 lines), road net (1687 lines), rail net (805 lines) and water net (716 lines) can be found. The objects on each layer have been stored as tuples of relational tables including also 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 (WARDS_URBAN_AREAS, WARDS_GREEN_AREAS, WARDS_ROADS, WARDS_RAILS and WARDS_WATERS) expressing non-disjointing relations.

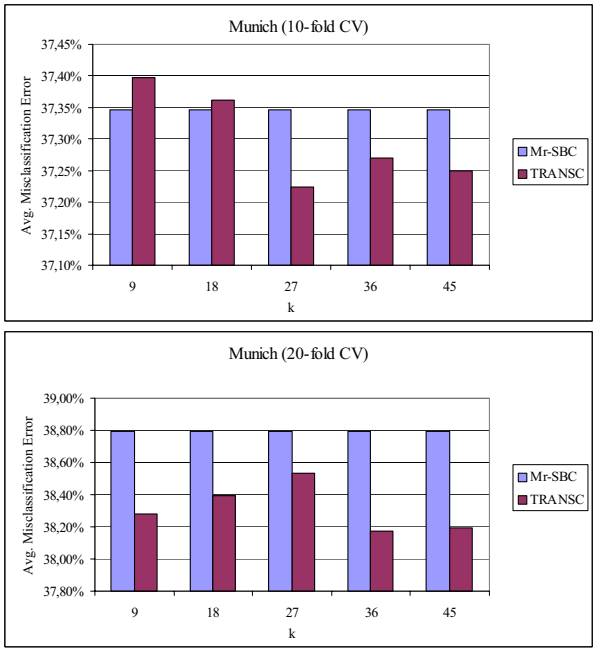

Fig. 3. TRANSC vs. Mr-SBC on Munich census data: average misclassification error on the working sets for 10-fold and 20-fold cross-validation

The number of materialized "non disjoint" relationships is 5313 (381 wards-urban areas, 13 wards-green areas, 2798 wards-roads, 1054 wards-rails and 1067 wards-waters).

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/mic Files/munich_db.tar.gz). The data have been collected in 1998 by Infratest Sozialforschung to develop the 1999 Munich rental guide. This dataset contains 2180 geo-referenced flats situated 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 as relational tables (CLOSE_TO and INSIDE).

The average misclassification error of TRANSC and Mr-SBC on both NWE Census Data and Munich Census Data is reported in Figure 2 and Figure 3, respectively. The reported results refer to both a 10-fold cross validation (CV) of the data and 20-fold cross validation of the same data. When experimenting on the NWE Census Data, we set $k \in \{2, 4, 7, 9, 1, 14\}$, while when experimenting on the Munich Census Data we set $k \in \{9, 18, 27, 36, 45\}$.

The results of Wilcoxon test are reported in Table 3 for the NWE Census Data and in Table 4 for the Munich Census Data. The results showed a slight improvement in the predictive accuracy of the transductive classifier over its inductive counterpart. Considering that both datasets are characterized by a strongly relevant structural component, these results confirm what observed with the Mutagenesis dataset, that is, the transductive approach we are proposing is beneficial when structural information is strongly relevant for the task at hand.

Table 3. TRANSC vs. Mr-SBC on NWE census data: results of the Wilcoxon test. Statistically significant p-values (< 0.05) are in italics. The sign + (-) indicates that TRANSC outperforms Mr-SBC (or vice-versa).

Experiment/k	2	4	6	8	10	12
10-fold CV	0.43 (+)	0.84 (+)	0.31 (+)	0.29 (+)	0.21 (+)	0.37 (+)
20-fold CV	0.12 (+)	0.17 (+)	0.36 (+)	0.12 (+)	0.09 (+)	0.16 (+)

Table 4. TRANSC vs. Mr-SBC on Munich census data: results of the Wilcoxon test. Statistically significant p-values (< 0.05) are in italics. The sign + (-) indicates that TRANSC outperforms Mr-SBC (or vice-versa).

Experiment/k	9	18	27	36	45
10-fold CV	0.42 (-)	0.74 (-)	*0.04* (+)	0.25 (+)	0.20 (+)
20-fold CV	*0.0019* (+)	*0.03* (+)	0.1 (+)	*0.00012* (+)	*0.00006* (+)

5 Conclusions

In this work we have investigated the combination of transductive inference with principled probabilistic MRDM classification in order to face the challenges posed by real-world applications characterized by both complex and heterogeneous data, which are naturally modeled as several tables of a relational database, and the availability of a small (large) set of labeled (unlabeled) data. Our proposed algorithm builds on an initial inductive classifier, namely a multi-relational naïve Bayesian classifier (Mr-SBC), learned from the training

(i.e., labeled) examples and used to perform a preliminary labeling of the working (i.e., unlabeled) data. The initial classification of the examples comprising the working set is then refined iteratively over a finite number of steps, each of which consists in a k-NN classification of all unlabeled examples and a subsequent reclassification of some "borderline" unlabeled examples. Neighbors are determined by computing a distance measure on a propositionalized representation of working examples. Propositionalization is based on the set of multi-relational rules mined by Mr-SBC.

The proposed transductive multi-relational classifier (TRANSC) has been compared to its inductive counterpart (Mr-SBC) in an empirical study involving both a benchmark relational dataset and two spatial datasets. The results of the experiments conducted on the benchmark dataset are in favor of TRANSC only when no background knowledge is considered (setting BK_0). Experimental results on spatial data are generally in favor of TRANSC and statistically significant in the case of the largest disproportion between training and working set (Munich census data with 20-fold cross validation). However, the improvements over the inductive counterpart are small. This findings confirm for the relational framework what already established for the propositional case [14], where similar small improvements have been observed when comparing SVMs in the inductive and transductive setting (SVMs vs TSVMs). Nonetheless, we intend to perfect our work in order to corroborate our intuition that transductive inference has benefits over inductive inference when applied to situations, like text mining, where the unlabeled examples heavily outnumber the labeled ones.

Acknowledgment

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

References

1. Bennett, K.P.: Combining support vector and mathematical programming methods for classification. pp. 307–326 (1999)
2. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: Proceedings of 18th International Conf. on Machine Learning, pp. 19–26. Morgan Kaufmann, San Francisco (2001)
3. Bosnic, Z., Kononenko, I., Robnic-Sikonja, M., Kukar, M.: Evaluation of prediction reliability in regression using the transduction principle. In: The IEEE Region 8 EUROCON 2003, pp. 99–103. IEEE Computer Society Press, Los Alamitos (2003)
4. Ceci, M., Appice, A.: Spatial associative classification: propositional vs structural approach. Journal of Intelligent Information Systems 27(3), 191–213 (2006)
5. Ceci, M., Appice, A., Malerba, D.: Mr-SBC: a multi-relational naive bayes classifier. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) PKDD 2003. LNCS (LNAI), vol. 2838, pp. 95–106. Springer, Heidelberg (2003)
6. Chen, Y., Wang, G., Dong, S.: Learning with progressive transductive support vector machines. Pattern Recognition Letters 24, 1845–1855 (2003)

7. De Raedt, L.: Attribute-value learning versus inductive logic programming: the missing links. In: Page, D. (ed.) Inductive Logic Programming. LNCS, vol. 1446, pp. 1–8. Springer, Heidelberg (1998)
8. Domingos, P., Pazzani, M.: On the optimality of the simple bayesian classifier under zeo-ones loss. Machine Learning 28(2-3), 103–130 (1997)
9. Flach, P.A., Lachiche, N.: Naive bayesian classification of structured data. Machine Learning 57(3), 233–269 (2004)
10. Gammerman, A., Azoury, K., Vapnik, V.: Learning by transduction. In: UAI 1998. Proc. of the 14th Annual Conference on Uncertainty in Artificial Intelligence, pp. 148–155. Morgan Kaufmann, San Francisco (1998)
11. Getoor, L.: Multi-relational data mining using probabilistic relational models: research summary. In: Knobbe, A., Van der Wallen, D.M.G. (eds.) Proc.of the 1st Workshop in Multi-relational Data Mining, Freiburg, Germany (2001)
12. Gora, G., Wojna, A.: RIONA: A classifier combining rule induction and k-nn method with automated selection of optimal neighbourhood. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, pp. 111–123. Springer, Heidelberg (2002)
13. Joachims, T.: Transductive inference for text classification using support vector machines. In: ICML 1999. Proc. of the 16th International Conference on Machine Learning, pp. 200–209. Morgan Kaufmann, San Francisco (1999)
14. Joachims, T.: Transductive learning via spectral graph partitioning. In: ICML 2003. Proc. of the 20th International Conference on Machine Learning, Morgan Kaufmann, San Francisco (2003)
15. Kukar, M., Kononenko, I.: Reliable classifications with machine learning. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.) ECML 2002. LNCS (LNAI), vol. 2430, pp. 219–231. Springer, Heidelberg (2002)
16. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.M.: Text classification from labeled and unlabeled documents using EM. Machine Learning 39(2/3), 103–134 (2000)
17. Orkin, M., Drogin, R.: Vital Statistics. McGraw Hill, New York (1990)
18. Pompe, U., Kononenko, I.: Naive bayesian classifier within *ilpr*. In Raedt, L.D. (ed) Proc. of the 5th Int. Workshop on Inductive Logic Programming, Dept. of Computer Science, Katholieke Universiteit Leuven, pp. 417–436 (1995)
19. Srinivasan, A., King, R.D., Muggleton, S.: The role of background knowledge: using a problem from chemistry to examine the performance of an ILP program. In Technical Report PRG-TR-08-99, Oxford University Computing Laboratory, Oxford (1999)
20. Vapnik, V.: Statistical Learning Theory. Wiley, Chichester (1998)
21. Wetteschereck, D.: A study of Distance-Based Machine Learning Algorithms. PhD thesis, PhD thesis, Department of Computer Science, Oregon State University, Corvalis, OR (1994)
22. Wrobel, S.: Relational Data Mining. In: chapter Inductive logic programming for knowledge discovery in databases. LNCS (LNAI), pp. 74–101. Springer, Heidelberg (2001)