TRANS-SC: a Transductive Structural Classifier

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

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

Abstract. The goal of transductive inference (or transduction) is to mine both labeled and unlabeled data within the same learning step and output a classification of the unlabeled examples with few errors as possible. We propose a new transductive classifier, named TRANS-SC, that works in a transductive setting and resorts to a principled probabilistic classification in multi-relational data mining to deal with structured data modeled as many tables as the number of object types. The method has been evaluated on some real-world relational data collections.

1 Introduction

Transductive classification provides a mechanism for mining "from the particular to the particular", without going through general models. The practical situation is that not only the labeled examples (training data) are available when mining a classifier, but also the unlabeled examples (working data) to be predicted. This suggests the idea of exploiting information coming from the examples to be predicted witih the mining step in order to output a classification of the unlabeled examples with few errors as possible and avoid to solve a more general mining problem as an intermediate step.

Despite transduction has been deeply investigated in the last decade, this research area is still in its infancy notwithstanding its great potential of application. All transductive algorithms proposed in literature assume data to be mined are represented in a single table (or database relation). In complex applications, this assumption turns out to be a great limitation, since data are heterogeneous and modeled as many tables as the number of object types.

In this paper, we present a novel transductive classifier, namely TRANS-SC (TRANsductive Sructural Classifier), that exploits the expressive power of Multi-Relational Data Mining (MRDM) to deal with relational data in its original structure and combines transductive inference with an improved version of the principled probabilistic MRDM classification presented in [1]. Probabilistic inference in this case allows to compute the class probability by returning not only the result of classification, but also the confidence of classification. This provides information on the potential uncertainty of classification.

The rest of the paper is organized as follows. In the next Section, we illustrate the probabilistic transduction in TRANS-SC, while in Section 3 we describe its application on some relational datasets.

2 Probabilistic Transduction in TRANS-SC

In the transductive inference setting, the relational classification problem can be formalized as follows: *Given*: a database schema S including a set of h relational tables $\{T_1, \ldots, T_h\}$; a set PK of primary key constraints on the tables in S; 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 that is different from the primary key of T and whose domain is the finite set $\{C_1, C_2, \ldots, C_L\}$; a training set that is an instance TS of the database schema S; a working set that is an instance WS of the database schema S with possibly unknown values for y; *Find* a prediction of the value of y for each example E of WS that is represented as a tuple $t \in WS.T$ and all tuples related to t in WS according to FK.

TRANS-SC classifies all examples of WS by extending the distance-weighted k-NN algorithm [2] and taking into account both training data and working data in the learning step. The idea is to classify each example $E \in WS$ on the basis of a k-sized neighborhood $N_k(E) = \{E_1, \ldots, E_k\}$ that is the set of the k examples $E_j \in WS$ closest to E according to a distance measure d. TRANS-SC returns 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 $y_i(E) = P(class(E) = C_i)$. Intuitively, the probability $P(class(E) = C_i)$ can be estimated as:

$$P(class(E) = C_i) = \frac{\#\{E_j \in N_k(E) | \hat{c}(E_j) = C_i\}}{k} \tag{1}$$

where $P(class(E) = C_i) \geq 0$ $(\forall i = 1 \ldots L)$ and $\sum_{i=1,\ldots,L} P(class(E) = C_i) = 1$. Equation 1 takes into account the class label $\hat{c}(E_j)$ predicted with an initial classifier $\hat{c}$ for each working example $E_j \in N_k(E)$. The classifier $\hat{c}$ is mined from TS by resorting to the MRDM probabilistic learning algorithm Mr-SBC [1] that is extended to take into account the cyclic paths on the relational data schema of S.

By looking at Equation 1 in deep, we note that all the k nearest neighbors of E equivalently contribute to estimate the class probability vector y'. In alternative, we propose to weight the contribution of each neighbor E_j according to the inverse of its distance from E so giving greater weight to closer neighbors. Let $w_j = \frac{1}{d(E,E_j)}$ be the weight associated to the neighbor $E_j \in N_k(E)$, we define $W_i(E) = \sum_{E_j \in N_k(E)} w_j \delta(C_i, \hat{c}(E_j))$, where $\delta(C_i, \hat{c}(E_j)) = 1$ if $\hat{c}(E_j) = C_i$, 0 otherwise. Hence, the output class probabilities is estimated as follows:

$$P(class(E) = C_i) = \frac{\frac{\#(\{E_j \in N_k(E) | \hat{c}(E_j) = C_i\})}{k} W_i(E)}{\sum_{t=1}^{L} \frac{\#(\{E_j \in N_k(E) | \hat{c}(E_j) = C_t\})}{k} W_t(E)} \tag{2}$$

The normalization by the summarization of all weighted class probabilities is required to guarantee that each probability class value is between 0 and 1 and the sum of the probabilities of all possible outcomes is 1. Finally, the single-class output $class(E)$ is obtained by returning the class C_i such that $P(class(E) = C_i) = max(P(class(E) = C_1), \ldots, P(class(E) = C_L))$. This weight-based estimation of class probabilities poses some problems in presence of neighbors $E_j \in N_k(E)$ with $d(E, E_j) = 0$ which lead to indeterminate form in Equation 2. To face this problem, we resort to the solution discussed in [2], that is, when there is one or more neighbors $E_j \in N_k(E)$ with $d(E, E_j) = 0$, the majority classification among them is assigned to E.

The Mr-SBC classifier $\hat{c}$ mined from TS is also involved in the computation of the distance d. Mr-SBC makes use of first-order classification rules to describe an example to be classified [1]. Let $\Re = \{A_j \Rightarrow y(X, C_i)\}$ be the set of classification rules extracted by Mr-SBC, whose consequent contains only one literal, that is, the class label while the antecedent contains a conjunction of at most MAX_LEN_PATH literals. Each example $E \in WS$ can be 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 $false$. Hence, TRANS-SC computes the distance between the examples E_1 and E_2 by resorting to a variant of the Hamming distance computed on V_{E_1} and V_{E_2}, that is, $d(E_1, E_2) = 1 - \frac{cardinality(V_{E_1}\ AND\ V_{E_2})}{cardinality(V_{E_1}\ OR\ V_{E_2})}$, where $cardinality(V)$ returns the number of $true$ values in the boolean vector V.

3 Empirical study

The empirical evaluation is done on North West England (NWE) Census data and Munich Census data by comparing TRANS-SC and Mr-SBC on the same 10-fold cross validation of data. Ten databases are created by randomly partitioning original ones and for each trial, both TRANS-SC and Mr-SBC are trained on a single database, while the hold-out nine databases form the working set. The non-parametric Wilcoxon test is used for the pairwise comparison of methods.

NWE data are provided in the European project SPIN! (http://www.ais.fraunhofer.de/KD/SPIN/project.html). These data involve the mortality rate (low, high) and continuous valued deprivation indexes available at level of 214 Greater Manchester census wards. The goal is to predict mortality rate by exploiting both deprivation factors and geographical factors concerning urban areas, green areas, road nets, rail nets and water nets overlapping Greater Manchester zone. Topological ("non disjoint") relationships between wards and objects in all these layers are materialized in relational tables. Differently, Munich Data (http://www.di.uniba.it/c̃eci/mic Files/munich_db.tar.gz) concern the degree (low, high) of monthly rent per square meter of 2180 flats geo-referenced within the Munich metropolitan area. this area is divided into three areal zones, each of which is decomposed into 64 districts, for a total of 446 subquarters. Subquarters and public transport stops within Munich are stored in relational tables. The phenomenon under observation is the "monthly rent per square meter". The spatial arrangement of data is expressed by both the "close_to" relation between metropolitan subquarters areas and the "inside" relation between public train stops and metropolitan subquarters.

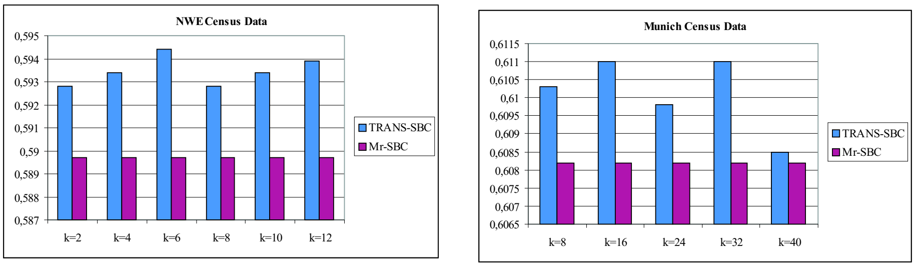

Fig. 1. TRANS-SC vs. Mr-SBC: 10 folds CV average accuracy on working sets.

Table 1. 10-fold CV results of the Wilcoxon test (p-value) by comparing average accuracy of TRANS-SC varying k vs. Mr-SBC. MAX_LENGTH_PATH=5.

Dataset	NWE Census Data						Munich Census Data				
k	2	4	6	8	10	12	8	16	24	32	40
p(sign)	.37(+)	.43(+)	.43(+)	.62(+)	.62(+)	.62(+)	.08(+)	.10(+)	.13(+)	.16(+)	.65(+)

The average accuracy of TRANS-SC and Mr-SBC is reported in Figure 1, while the results of the Wilcoxon test are reported in Table 1. The sign "+" indicates TRANS-SC outperforms Mr-SBC at p value of significance. Results show that classification takes advantage from the transductive framework independently from k value. These results confirm improved accuracy of the transductive classifier. The observed slight improvement is explained with the inherent complexity of the tasks. Moreover, accuracy improvement is statistically significant when increasing the working set size. This is obtained by evaluating TRANS-SC and Mr-SBC on a 20-fold cross validation of Munich Census Data that is the largest data collection involved in this empirical study. In this case, results of Wilcoxon test are always statistically in favor of TRAN-SC with $p \leq 0.09$. To corroborate our intuition on benefits of transductive inference, further applications involving "very large" unlabeled data collections should be considered.

Acknowledgment

This work is partial fulfillment of the research objective of ATENEO-2005 project "Gestione dell'informazione non strutturata: modelli, metodi e architetture".

References

1. M. Ceci, A. Appice, and D. Malerba. Mr-SBC: a multi-relational naive bayes classifier. In *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD 2003*, volume 2838 of *LNAI*, pages 95–106. Springer-Verlag, 2003.
2. T. Mitchell. *Machine Learning*. McGraw Hill, New York, USA, 1997.