FEATURE EXTRACTION FOR DYNAMIC INTEGRATION OF CLASSIFIERS IN MEDICAL PROBLEMS MODELLING

Alexey Tsymbal[1], Mykola Pechenizkiy[2], Seppo Puuronen[2]

[1] Trinity College Dublin, Department of Computer Science, Dublin, Ireland

[2] University of Jyväskylä, Dept. of CS and IS, Jyväskylä, Finland

Abstract – The goal of this paper is to introduce and examine three feature extraction techniques for the dynamic integration of classifiers with regard to their application to medical problems modelling. In this paper, we evaluate accuracy of FEDIC algorithm for the problem of the modelling of acute abdominal pain, and Liver Disorders data set from the UCI machine learning repository.

Introduction. Current electronic data repositories, especially in medical domains, contain enormous amounts of data. These data include also currently unknown and potentially interesting patterns and relations that can be uncovered using knowledge discovery and data mining methods. Inductive learning systems were successfully applied in a number of medical domains, e.g. in the localization of a primary tumor, prognostics of recurrence of breast cancer, diagnosis of thyroid diseases, and rheumatology.

Numerous data mining methods have recently been developed to extract knowledge from these large databases. Selection of the most appropriate data-mining method or a group of the most appropriate methods is usually not straightforward. Often the method selection is done statically for all new instances of the domain area without analyzing each particular new instance. Usually better data mining results can be achieved if the method selection is done dynamically taking into account characteristics of each new instance.

Recent research has proved the benefits of the use of ensembles of classifiers for classification problems [2]. An ensemble is often more accurate than any of the single classifiers in the ensemble. The ensemble approach entails two essential questions: (1) which classifiers to use as the components of the ensemble (generation of the base

classifiers) and (2) how to combine their individual predictions into a single final classification (the integration procedure).

Both theoretical and empirical research have demonstrated that a good ensemble is one where the base classifiers in the ensemble are both accurate and tend to err in different parts of the input space (e.g., have high diversity in their predictions). One efficient way to construct an ensemble of diverse classifiers is to use different feature subsets. The second issue in creating an effective ensemble is the choice of the function for combining the predictions of the base classifiers. It was shown that increasing coverage of an ensemble through diversity is not enough to insure increased prediction accuracy – if the integration method does not utilize coverage, then no benefit arises from integrating multiple models.

In many real-world applications, numerous features are used in an attempt to ensure accurate classification. If all those features are used to build up classifiers, then they operate in high dimensions, and the learning process becomes analytically and computationally complicated. For instance, many classification techniques are based on Bayes decision theory or on nearest neighbor search, which suffer from the so-called "curse of dimensionality" due to the drastic rise of computational complexity and classification error in high dimensions [3]. Hence, there is a need to reduce the dimensionality of the feature space before classification.

Feature extraction is a dimensionality reduction technique that extracts a subset of new features from the original set of features by means of some functional mapping keeping as much information in the data as possible [3].

Dynamic Integration of Classifiers with Instance Space Transformation. In this paper, we consider the use of feature extraction in order to cope with the curse of dimensionality in the dynamic integration of classifiers. We propose the FEDIC (Feature Extraction for Dynamic Integration of Classifiers) algorithm, which combines the dynamic selection, dynamic voting and dynamic voting with selection integration techniques (DS, DV and DVS) with the conventional Principal Component Analysis (PCA) and two supervised eigenvector-based feature extraction approaches (that use the within- and between-class covariance matrices). The first

eigenvector-based approach is parametric, and the other one is nonparametric. Both these take class information into account when extracting features, in contrast to PCA [3].

The FEDIC learning model that was introduced in [4] consists of five phases: (1) the training of the base classifiers phase; (2) the feature extraction phase (FE); (3) the dynamic integration phase (DIC); (4) the model validation phase; and (5) the model testing phase. The model is built using a wrapper approach, where the variable parameters in FE and DIC can be adjusted to improve performance as measured at the model validation phase in an iterative manner.

Experiments. The experiments are conducted on three large data sets with the cases of acute abdominal pain (AAP): (1) Small-AAP I, (2) Medium-AAP II, and (3) Large-AAP III, with the numbers of instances equal correspondingly to 1254, 2286, and 4020 and the BUPA Liver Disorders data set from the UCI machine learning repository [1]. AAP data sets represent the same problem of separating acute appendicitis (class "appendicitis"), which is a special problem of acute abdominal pain, from other diseases that cause acute abdominal pain (class "other diagnoses"). The early and accurate diagnosis of acute appendicitis is still a difficult and challenging problem in everyday clinical routine.

To construct the ensembles of classifiers we have used the EFS_SBC (Ensemble Feature Selection for the Simple Bayesian Classification) algorithm. Experiment design was done as in [4].

Results and Discussions. From Table 1, one can see that for the Large APP III data set every approach gives almost the same accuracy. On the Medium APP II data set static integration (bestSIC) is better than the simple global Bayesian classifier (Bayes); dynamic integration (BestDIC) shows better results than the static one; dynamic integration in the space of transformed features (BestFEDIC) does not influence on the accuracy results significantly (the statistical significance is checked with the 1-tailed Student t-test with 0.95 level of significance). On the Small APP I data set the situation is completely different: simple Bayes, static selection and dynamic selection show almost the same accuracy results, while feature extraction for

dynamic integration significantly improves the accuracy.

The mean of sensitivity and specificity (or average class accuracy) on the Small APP I data set rivals the best previously published results for this data set [20].

Table 1 – Classification accuracies for AAP data sets

	Large AAP III	Medium AAP II	Small AAP I
Bayes	0.847	0.523	0.769
bestSIC	0.848	0.544	0.771
bestDIC	0.848	0.566	0.772
bestFEDIC	0.848	0.558	**0.782**

References

[1] C.L. Blake, C.J. Merz, UCI repository of machine learning databases [http://www.ics.uci.edu/~mlearn/ MLRepository.html]. Dept. of Information and Computer Science, University of California, Irvine, CA, 1998.

[2] T. G. Dietterich, "Ensemble Learning Methods", In: M.A. Arbib (ed.), Handbook of Brain Theory and *Neural Networks*, 2nd ed., MIT Press, 2001.

[3] K. Fukunaga, Introduction to statistical pattern recognition. Academic Press, London, 1999.

[4] Tsymbal A., Pechenizkiy M., Puuronen S., Patterson D.W., Dynamic integration of classifiers in the space of principal components, In: L.Kalinichenko, R.Manthey, B.Thalheim, U.Wloka (eds.), Proc. Advances in Databases and Information Systems: 7th East-European Conf. ADBIS'03, Dresden, Germany, Lecture Notes in Computer Science, Vol. 2798, Springer-Verlag, 2003, 278-292.