PRE-PROCESSING OF METEOROLOGICAL DATA IN KNOWLEDGE DISCOVERY

E. TSAGALIDIS[1], G. EVANGELIDIS[2]

[1]*Hellenic Agricultural Insurance Organization, Meteorological Applications Centre*
[2]*Department of Applied Informatics, University of Macedonia*

ABSTRACT

As an application of Knowledge Discovery from Meteorological Databases, we attempt to relate recorded precipitation data from the Mikra Meteorological Station in Thessaloniki, Greece, to the recorded data of the closest to the station ERA-40 node (i.e., node with latitude 40^0N and longitude 22.5^0E). We use the daily values of the ERA-40 meteorological data from the European Centre for Medium-Range Weather Forecasts (four times a day at 00h, 06h, 12h and 18h) for a period of 42 years (1960-2001). We examine different scenarios in the pre-processing phase of the input data and we study their impact on the performance of data mining algorithms designed to predict the occurrence of precipitation in Mikra Meteorological Station. The goal is to determine the appropriate pre-processing on the input file that can ensure effective application of data mining techniques.

ΕΠΕΞΕΡΓΑΣΙΑ ΜΕΤΕΩΡΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΕΙΣΟΔΟΥ ΣΕ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΓΝΩΣΗΣ

Ε. ΤΣΑΓΚΑΛΙΔΗΣ[1], Γ. ΕΥΑΓΓΕΛΙΔΗΣ[2]

[1]*Οργανισμός Ελληνικών Γεωργικών Ασφαλίσεων, Κέντρο Μετεωρολογικών Εφαρμογών*
[2]*Τμήμα Εφαρμοσμένης Πληροφορικής, Πανεπιστήμιο Μακεδονίας.*

ΠΕΡΙΛΗΨΗ

Χρησιμοποιούνται οι ημερήσιες τιμές των μετεωρολογικών παραμέτρων των δεδομένων ERA-40 του Ευρωπαϊκού Κέντρου Μεσοπρόθεσμων Προγνώσεων Καιρού (τέσσερις φορές την ημέρα στις 00h, 06h, 12h και 18h), του πλησιέστερου κόμβου στο Μετεωρολογικό Σταθμό Μίκρας Θεσσαλονίκης για μια περίοδο 42 ετών (1960-2001). Επιχειρείται η συσχέτιση των δεδομένων του κομβικού σημείου με γεωγραφικό πλάτος 40^0Β και μήκος 22.5^0Α, με την καταγραφή υετού στο Μ.Σ Μίκρας, ως μια εφαρμογή της Ανακάλυψης Γνώσης από Μετεωρολογικές Βάσεις Δεδομένων. Εξετάζονται διαφορετικά σενάρια στο στάδιο προ-επεξεργασίας των δεδομένων εισόδου και μελετάται η επίδρασή τους στην απόδοση Αλγορίθμων Εξόρυξης Γνώσης που σκοπό έχουν την πρόβλεψη της εμφάνισης υετού στο Μ.Σ Μίκρας. Στόχος είναι να υποδειχθεί η μέθοδος διαμόρφωσης του αρχείου εισόδου που μπορεί να εξασφαλίσει την αποτελεσματικότερη εφαρμογή τεχνικών Εξόρυξης Γνώσης.

1. INTRODUCTION

"Knowledge Discovery in Databases (KDD) is an interactive, iterative procedure that attempts to extract implicit, previously unknown, and potentially useful knowledge from data" (Roiger et al., 2003). Once a specific problem has been defined, an

appropriate dataset is chosen and goes through a pre-processing and transformation phase. The resulting dataset is analyzed by applying various data mining algorithms on it. The output is interpreted and evaluated and decisions are made about whether to repeat previous steps or take action incorporating and applying the extracted knowledge directly to appropriate problems (Roiger et al., 2003).

Knowledge Discovery has become a very popular scientific discovery tool during the past decade. Problems from diverse application fields, like, Astronomy, Athletics, Meteorology, Marketing, etc., can be seen as interesting data mining applications that involve interdisciplinary collaboration.

In this paper we address the problem of data preparation in the context of the application of data mining algorithms on Meteorological Databases. We combine two datasets, the ERA-40 data from the reanalysis project of the European Centre for Medium-Range Weather Forecasts (ECMWF), and the weather observations data from the Meteorological Station of Mikra (Thessaloniki, Greece). Our goal is to predict the occurrence of precipitation in the station. This can be addressed as a classification problem, i.e., how to assign a set of ERA-40 parameter values to one of a set of well-defined classes. More specifically, we are interested in the presence or absence of precipitation on the ground (i.e., we have a binary class variable). We apply five different data mining algorithms to design models for predicting our class variable. The model evaluation is performed using a training/test set method, where the output consists of the evaluation metrics.

Experiments have shown that redundant attributes as input variables are responsible for significant losses of performance in standard classifiers (Roiger et al., 2003). Input attributes highly correlated with other input attributes are redundant, as it happens in the ERA-40 dataset. We address this issue by using the Principal Component Analysis (PCA) extraction method as a data reduction technique on our original data.

A second question that we address is whether data transformation can improve the performance of data mining algorithms. We experiment with three different scenarios for preparing the dataset as input to the algorithms in order to find the most appropriate for the ERA-40 data. In addition, when classes are imbalanced, many learning algorithms can suffer from the perspective of reduced performance (Jo et al., 2004). In our problem, the positive class (recorded precipitation) is only 16.1% of the cases. As an attempt to overcome this issue of skewed data, we apply the common solution of sampling the data randomly to achieve a balanced distribution (Jo et al., 2004; Weiss, 2004; Batista et al., 2004).

The remainder of the paper is organized as follows. Section 2 describes the datasets we used for applying the data mining algorithms. Section 3 discusses the methodology used in the experiments. In Section 4 we present the analysis and the results, and, finally, we conclude in Section 5.

2. DATASETS

The European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis ERA-40 is a global atmospheric analysis of many conventional observations and satellite data streams for the period September 1957 to August 2002. Over the past decade, reanalyses of multi-decadal series of past observations have become an important and widely utilized resource for the study of atmospheric and oceanic processes and predictability (ECMWF). Since reanalyses are produced using fixed,

modern versions of the data assimilation systems developed for numerical weather prediction, they are more suitable than operational analyses for use in studies of long-term variability in climate. Reanalysis products are used increasingly in many fields that require an observational record of the state of either the atmosphere or its underlying land and ocean surfaces. The main objective of the reanalysis project ERA-40 is to promote the use of global analyses of the state of the atmosphere, land and surface conditions over the period 1957-2002 (ECMWF).

There are numerous data products that are separated into dataset series based on resolution, vertical coordinate reference, and likely research applications. In this study, we used the ERA-40 2.5 degree latitude-longitude gridded upper air analysis on pressure surfaces. This dataset contains 11 variables on 23 pressure surfaces on an equally spaced global 2.5 degree latitude-longitude grid. All variables are reported four times a day at 00, 06, 12 and 18UTC for the entire period.

We created our initial dataset choosing the values of 10 variables on 7 pressure surfaces on one node. We used only the data from node with geographical coordinates 40^0N latitude and 22.5^0E longitude, which is the closest node to the Meteorological Station of Mikra, Thessaloniki, Greece. The 10 variables are the *geopotential* in $m^2 \cdot s^{-2}$, *temperature* in K, *U velocity* in $m \cdot s^{-1}$, *V velocity* in $m \cdot s^{-1}$, *specific humidity* in $kg \cdot kg^{-1}$, *relative humitidy* as percentage (%), *vorticity (relative)* in s^{-1}, *potential vorticity* in $K \cdot m^2 \cdot kg^{-1} \cdot s^{-1}$, *divergence* in s^{-1}, and *vertical velocity* in $Pa \cdot s^{-1}$. We omit the *11th Ozone mass mixing ratio*. The 1000hPa, 925hPa, 850hPa, 700hPa, 500hPa, 300hPa and 200hPa are the 7 pressure surfaces we chose, because these are the ones that are mainly used by the meteorology forecasters operationally. In addition, the values of the *barometric pressure on mean sea level* in Pa, supplement the initial dataset that consists of 71 variables.

Furthermore, the initial values of most of the variables for each pressure surface and the pressure on mean sea level were transformed to make them easier to understand. More specifically, *specific humidity* was converted to $g \cdot kg^{-1}$ and *vertical velocity* to $hPa \cdot h^{-1}$. The relatively small values of both *vorticity (relative)* and *divergence* were multiplied by 10^6, and the value of *potential vorticity* by 10^8. Regarding the wind, *wind direction* in azimuth degrees and *wind speed* in knots were calculated using the U and V velocities. Also, the azimuth degrees for the *wind direction* were assigned into the eight discrete values of north (N), northeast (NE), etc., used in meteorology. The *geopotential* was divided by the World Meteorological Organization (WMO) defined gravity constant of $9.80665 m \cdot s^{-2}$, thus, it was transformed to *geopotential height* in gpm. Finally, the values of *barometric pressure on mean sea level* were expressed in hPa, and only the values of *temperature* and *relative humidity* on pressure surfaces remained unchanged.

The 6-hourly main synoptic surface observation data of the Mikra Meteorological Station, located at 40.52^0N, 22.97^0E and altitude of 4m, completed our initial dataset. More specifically, we collected the recorded precipitation data of the period 1/1/1960 00UTC – 31/12/2001 18UTC. We assigned the value 'yes' to the 6-hourly records of rain, drizzle, sleet, snow, shower at the station or the records of thunderstorm at the station or around it, and the value 'no' to the rest of the records, thus, creating the class variable of our study. We mention that the determination of the recorded precipitation is taking into account both the present and past weather of the synoptic observation and that snow or thunder have priority over rain. Table 1 depicts the distribution of the

precipitation types that had been recorded in the Mikra Meteorological Station according to the defined subclusters.

TABLE 1. Natural distribution of values within the precipitation class variable

Precipitation 'yes'				Precipitation 'no'		
Rain, Drizzle	Snow, Sleet	Thunder	**Total**	Fog	Fair, Cloudy	**Total**
11.66%	0.89%	3.55%	**16.1%**	2.27%	81.62%	**83.9%**

3. METHODOLOGY

We applied data reduction using the Principal Component Analysis (PCA) extraction method to remove highly correlated variables from the ERA-40 dataset. We used the SPSS statistical software package to process the entire ERA-40 dataset (SPSS). We applied PCA and we selected components with eigenvalues greater than one. Then, we examined the component matrix of loadings. Component loadings represent the degree of association, correlation, of each variable with each component. To identify the significant loadings for each variable, the lower threshold of the absolute value of 0.4 was used due to the relatively large dataset size. In addition, the communalities of the variables, which were calculated as the sum of the squared component loadings, were examined to identify variables that explain more than 50% of the variance in each variable.

Furthermore, we employed the Varimax with Kaiser normalization orthogonal rotation method to achieve simpler and more meaningful component solutions, reducing some of the ambiguities that often accompany initial unrotated solutions. When a variable is found to have no significant loadings or just one significant loading, the variable's communality is deemed to be too low or to have a cross-loading. Then, the variable is candidate for deletion and the model is respecified. A variable has a cross-loading when it has more than one significant loadings (Hair et al., 2006; SPSS).

Common operational practice for meteorologists in weather forecasting is the calculation of the differences between successive values of the meteorological variables. According to this practice and after the application of data reduction, we experimented with two extra scenarios in addition to the one that uses the regular values of the selected variables. In the second scenario, the values at each synoptic hour were replaced by the past 24-hourly differences, and, in the third scenario, in addition to the 24-hourly differences we kept the regular values of the variables that express the wind elements. We express the differences in the case of wind direction by using values -1, -2, -3 and -4 when the wind turns counter clock-wise, 1, 2 and 3 when it turns clock-wise, and 0 when the wind direction remains the same. For example, when a western wind turns to southwest the difference is –1, when it turns to south the difference is –2, when it turns to northwest the difference is 1, etc.

In a concept-learning problem, we have class imbalance in our data if some classes have a much larger number of instances than the rest. Such a situation poses challenges for typical classifiers, such as decision tree induction systems or multilayer perceptrons, since they are designed to optimize overall accuracy without taking into account the relative distribution of each class. As a result, these classifiers tend to ignore small classes while concentrating on accurately classifying the large ones. Such problems occur in a large number of practical domains and especially in our dataset, where only 16.1% of the instances correspond to precipitation. In an effort to improve the

performance of classifiers it is common to use the re-sampling process of manipulating the distribution of the training instances. More specifically, the random under-sampling method is applied to randomly remove instances in the majority class and to reduce them to the size of the minority class, thus, producing a completely balanced distribution (Jo et al., 2004; Weiss, 2004; Batista et al., 2004).

4. EXPERIMENTS AND RESULTS

4.1. Feature selection

After applying PCA and examining the component matrix of loadings and the variable communalities, we deleted a total of 36 variables from our initial dataset that consisted of 71 variables. The component model was respecified six times with a final outcome of 35 variables and 9 components with eigenvalues greater than 1. The analysis reveals that the first component is most highly correlated with the geopotential height on 200hPa, and, generally, is highly correlated with the geopotential height in the upper levels, the temperature almost in all levels, and the specific humidity in low levels of the atmosphere. The second component is most highly correlated with the relative vorticity on 1000hPa, and, generally, it is highly correlated with the relative vorticity in low levels, the geopotential height on 925hPa, and the pressure at mean sea level. The third component is highly correlated with the wind direction in middle and upper levels and especially on 300hPa. The fourth component is highly correlated with the wind speed in upper levels and especially on 300hPa. The fifth component is highly correlated with the wind speed in low levels and especially on 925hPa. The sixth component is most highly correlated with the divergence on 300hPa and also the vertical velocity in the upper levels. The seventh component is highly correlated with the temperature and relative vorticity on 200hPa. The eighth component is most highly correlated with the potential vorticity on 500hPa and also the relative vorticity in the same level. Finally, the ninth component is highly correlated with the wind direction in low levels and especially on 925hPa.

Table 2 displays the variance explained by the rotated components and additionally the corresponding nine most highly correlated variables. The *Total* column gives the eigenvalue, or amount of variance in the original variables accounted for by each component. The *% of Variance* column gives the ratio, expressed as a percentage, of the variance accounted for by each component to the total variance in all of the variables. The *Cumulative %* column gives the percentage of variance accounted for by the first 9 components (SPSS). They explain nearly 85.2% of the variability in the original variables and it is possible to considerably reduce the complexity of the data set by using these components, with a 14.8% loss of information. As a result, we can reduce the size of the ERA-40 dataset by selecting the 9 most highly correlated variables with the 9 principal components.

TABLE 2. Variance explained by rotated components and the representative variables

Component	Variable	Total	% of Variance	Cumulative %
1	geopotential height 200hPa	9.8	28.0	28.0
2	relative vorticity 1000hPa	4.2	11.9	39.9
3	wind direction 300hPa	2.9	8.4	48.3
4	wind speed 300hPa	2.6	7.5	55.7
5	wind speed 925hPa	2.4	7.0	62.7

6	divergence 300hPa	2.3	6.7	69.4
7	Temperature 200hPa	2.1	6.0	75.4
8	potential vorticity 500hPa	1.8	5.0	80.4
9	wind direction 925hPa	1.7	4.8	85.2

4.2. Data transformation

The reduced ERA-40 dataset with the 9 chosen variables, as predictors, and the precipitation, as class variable, comprise our experimental dataset (Dataset1). As explained in the previous section, we formed a second dataset by replacing the regular values of the 9 predictors with the past 24-hourly differences (Dataset2). We also formed a third dataset with 13 predictors by replacing the regular values of the 9 predictors with the past 24-hourly differences plus the regular values of the wind direction on 300hPa and 925hPa and the wind speed on 300hPa and 925hPa (Dataset3).

The three datasets were the input to five data mining algorithms that were run and evaluated using WEKA. The algorithms were the decision tree C4.5 with unpruning and Laplace estimate, the k-Nearest Neighbours with k=3 and Euclidean distance, the RIPPER, the Naïve Bayesian, and the Multilayer Perceptron neural network with backpropagation. The last three algorithms were run using the default settings of WEKA (Hall et al., 2009; Witten et al., 2005).

As an evaluation metric, we used the Area Under the ROC (Receiver Operating Characteristics) curve or simply AUC that measures the performance of the algorithms as a single scalar. ROC graphs are two-dimensional graphs in which the True Positive Rate (the percentage of positive cases correctly classified as belonging to the positive class) is plotted on the Y axis and the False Positive Rate (the percentage of negative cases misclassified as belonging to the positive class) is plotted on the X axis. A ROC graph depicts relative tradeoffs between benefits (true positives) and costs (false positives). The AUC is a reliable measure especially for imbalanced datasets to get a score for the general performance of a classifier and to compare it to that of another classifier (Weiss, 2004; Batista et al., 2004).

The training/test set method was used to build and evaluate the models. Each one of the three initial datasets (scenarios) with 61364 examples or instances was divided into 10 non-overlapping folds. By taking each one of the 10 folds as test set and the rest 9 as a pool of instances for choosing the training sets, we formed 10 groups with 55228 training instances and 6136 test instances. Then, from the training instances of each group, we randomly took 10 samples with replacement consisting of 17788 instances. Thus, we formed 100 training/test datasets with 23924 instances (17788 training and 6136 test instances, 74.35% - 25.65%) for each scenario for a total of 300 datasets. Every fold or sample was chosen randomly, but it followed the natural distribution according to the clusters within the precipitation class variable, as shown in Table 1.

Moreover, for each one of the 10 groups, we formed 10 balanced training sets, where the number of ‘yes’ was equal to the number of ‘no’. These datasets comprised of 8894 randomly with replication selected instances of ‘no’, and all the 8894 available instances of ‘yes’. We used the same 10 test folds that cover all the initial 61364 instances, to evaluate the built models. It is noted that every sample of the ‘no’ cluster of precipitation taken randomly, follows the natural distribution according to its subclusters as shown in Table 1.

To recap, we test each one of the three scenarios with 100 training/test datasets that follow the natural distribution and 100 training/test datasets that follow the balanced distribution, for a total of 600 training/test datasets. These datasets comprise the input to the five data mining algorithms. Thus, we performed 3000 runs in the WEKA environment and we show the results in Table 3 and Figure 1. Table 3 presents the mean value and the standard deviation of AUC of the 100 runs for each dataset, distribution and algorithm.

TABLE 3. Mean value and standard deviation of AUC

	Natural distribution						Balanced distribution					
	Dataset1		Dataset2		Dataset3		Dataset1		Dataset2		Dataset3	
Algorithm	$\overline{x}$	S.D.	$\overline{x}$	S.D.	$\overline{x}$	S.D.	$\overline{x}$	S.D.	$\overline{x}$	S.D.	$\overline{x}$	S.D.
D. Tree	.728	.011	.641	.009	.679	.009	.737	.008	.653	.009	.683	.010
k-NN	.679	.009	.588	.009	.639	.011	.734	.007	.623	.010	.693	.011
MPbp	.786	.009	.703	.008	.727	.011	.803	.008	.716	.005	.758	.009
N. Bayes	.773	.008	.703	.007	.736	.007	.774	.008	.704	.006	.737	.007
RIPPER	.586	.014	.538	.007	.546	.009	.732	.010	.656	.009	.684	.011

Figure 1 depicts the boxplots of the corresponding AUC values. The white boxplots correspond to Dataset1, the light grey to Dataset2, and the dark grey to Dataset3. The boxplots for the balanced datasets have a pattern consisting of black dots, whereas, the boxplots for the natural datasets have no pattern. We notice that the balanced distribution strategy performs better than the natural one, especially for the k-Nearest Neighbour, the RIPPER and the Multilayer Perceptron, whereas, it performs slightly better for the Decision Tree and almost the same for the Naïve Bayesian. Concerning the three different scenarios for each distribution and algorithm, Dataset1 (with the regular values) performs better than the other two, and, Dataset3 (with the differences plus the regular values) performs better than Dataset2 (with only the differences). Finally, in the natural distribution strategy the algorithm ranking is Multilayer Perceptron, Naïve Bayesian, Decision Tree, k- Nearest Neighbour and RIPPER. In the case of the balanced distribution strategy the performance of k- Nearest Neighbour and RIPPER vastly increases and it is almost equal to that of Decision Tree, whereas the performance of Multilayer Perceptron is the best one and of Naïve Bayesian the second best.

5. CONCLUSIONS

We applied Principal Component Analysis to reduce the 71 initial chosen variables of the ERA-40 dataset to 9 uncorrelated to each other variables. The variables that could represent the 9 principal components explaining nearly 85.2% of the variability in the original variables were the following: geopotential height on 200hPa, relative vorticity on 1000hPa, wind direction on 300hPa, wind speed on 300hPa and on 925hPa, divergence on 300hPa, temperature on 200hPa, potential vorticity on 500hPa and wind direction on 925hPa.

The reduced ERA-40 dataset and the historical precipitation records of the Meteorological Station of Mikra were the input into five data mining algorithms we

used to build models that predict the occurrence of precipitation at the station.

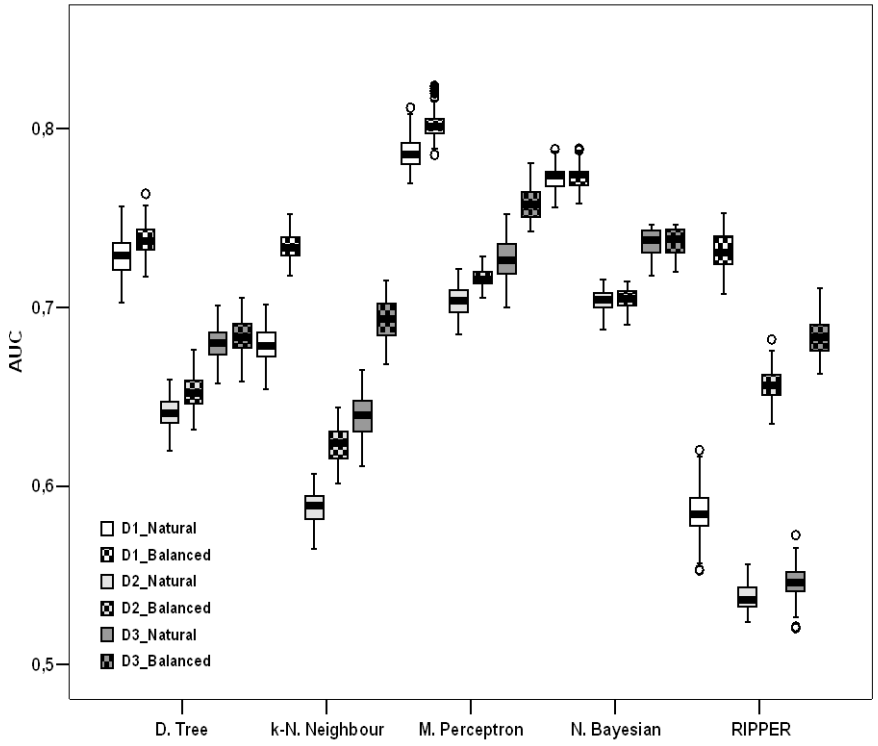

FIGURE 1. Boxplots of AUC values for each algorithm, dataset and distribution.

The dataset with the regular values of the ERA-40 variables (Dataset1) achieved better performance than the one with the past 24-hourly differences of the ERA-40 variables (Dataset2), and the one with the differences plus the regular values of the variables expressing the wind elements (Dataset3). In addition, when using balanced rather than natural distribution training sets, according to the values of precipitation, all algorithms perform better on all three datasets with the exception of the Naïve Bayesian algorithm. Finally, the Multilayer Perceptron neural network with backpropagation algorithm outperforms all other algorithms, revealing the most effective data mining algorithm in this meteorological domain.

ACKNOWLEDGEMENTS

We wish to thank the European Centre for Medium-Range Weather Forecasts and the Hellenic National Meteorological Service for providing us with the ERA-40 dataset and the historical meteorological data of the Mikra Meteorological Station, respectively.

REFERENCES

Batista G, Prati R, Monard MC, 2004: A study of the behaviour of several methods for balancing Machine Learning Training Data, *SIGKDD Explorations*, 6, Issue 1, 20-29.

European Centre for Medium-Range Weather Forecasts (ECMWF, ERA-40): http://www.ecmwf.int/research/era/do/get/era-40.

Hair J, Black W, Babin B, Anderson R, Tatham R, 2006: *Multivariate Data Analysis, 6th Edition.* Prentice Hall.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH, 2009: The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, 11(1).

Jo T, Japkowicz N, 2004: Class Imbalances versus Small Disjuncts, *SIGKDD Explorations*, 6, Issue 1, 40-49.

Roiger R, Geatz M, 2003: *Data Mining: A Tutorial-based Primer*. Addison-Wesley, 408 pp.

SPSS: *Statistical Analysis Software*. http://www.spss.com.

Weiss MG, 2004: Mining with Rarity: A Unifying Framework, *SIGKDD Explorations*, 6, Issue 1, 7-19.

Witten I, Frank E, 2005: *Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition.* Morgan Kaufmann, 525pp.