TASK QUARTERLY 11 No 1–2, 103–115

AN ALGORITHM FOR DATA QUALITY ASSESSMENT IN PREDICTIVE TOXICOLOGY

LADAN MALAZIZI[1], DANIEL NEAGU[1]
AND QASIM CHAUDHRY[2]

[1]*Department of Computing, School of Informatics, University of Bradford, Great Horton Road, Bradford, BD7 1DP, UK*
{l.malazizi, d.neagu}@bradford.ac.uk

[2]*Central Science Laboratory, Sand Hutton, York, YO41 1LZ*
q.chaudhry@csl.gov.uk

(Received 21 December 2006; revised manuscript received 26 January 2007)

Abstract: Lack of the quality of the information that is integrated from heterogeneous sources is an important issue in many scientific domains. In toxicology the importance is even greater since the data is used for Quantitative Structure Activity Relationship (QSAR) modeling for prediction of chemical toxicity of new compounds. Much work has been done on QSARs but little attention has been paid to the quality of the data used. The underlying concept points to the absence of the quality criteria framework in this domain. This paper presents a review on some of the existing data quality assessment methods in various domains and their relevance and possible application to predictive toxicology, highlights number of data quality deficiencies from experimental work on internal data and also proposes some quality metrics and an algorithm for assessing data quality concluded from the results.

Keywords: QSAR models, data quality, data cleaning

1. Introduction

Nowadays, given the development and low cost of high data storage capacities, more experimental data is available from various scientific laboratories. A modern approach to the accessibility of large amounts of data is therefore using data integration methods. In this context, data quality is one of the most important attributes for data integration. A special case is predictive toxicology, the science of developing in-silico models for toxicity prediction, which is of interest to chemical and pharmaceutical industry, regulatory bodies and environmental protection agencies. This approach considers the use of experimental data for Quantitative Structure Activity Relationship modeling [1], relating aspects of chemical compound structure to biological activities against various endpoints in order to predict chemical toxicity of new compounds. Currently, most of toxicity data is obtained from publicly available databases such as

Toxnet [2] or DSSTox [3] as collation of various experimental data from governmental or industrial bodies. But because of their limitations such as various experimental conditions, incomplete source identification or lack of standardization requirements for different measurement units, many of them may still not be fully recognized as reliable sources. Efforts are paid to organize and manage toxicity databases toward standardization and to improve their integrity and reliability by National Institute of Standards and Technology [4] which focuses on producing a common vocabulary of weights, measures, names and symbols to scientific enterprises and agreement of a data file terminologies. This effort provides procedural guidelines for experimental work but still the inconsistencies of data values within a source or from one source to another remain a subject to be addressed. These drawbacks generated a demand of methods to tackle the data quality problems [5].

In the next section, we overview some integration methods implemented in other domains to address the problem of low quality data. A short analysis of each method is also provided to clarify the relevance of each approach to the predictive toxicology domain. A summary of extensive experimental work carried out in our research laboratory on assessment of data quality in predictive toxicology domain using data for five different endpoints is presented in Section 3. Results of this work have been used to originate some criteria for measuring data quality and could define a foundation for future studies in automated data-driven model development and validation.

2. Examples of data quality assessment methods

There have been various methods developed to approach the problem of data quality assessment in different domains, depending on some specific criteria. These methods are mainly based on user, domain and use of the system constraints:

1. According to Naumann [6] Information Quality (IQ) depends on the user, the information and the process of accessing the information. For example, for user there are some information quality criteria such as: believability and concise representation. Criteria for the information itself include: completeness and customer support, assessed through parsing, sampling and expert input. For the process, quality criteria include: availability and accuracy assessed by cleansing techniques and parsing. The IQ method identifies elements that are information system processing oriented. Other information quality measurements such as believability are also user dependent assessment methods and could vary from one user to another. Process-based criteria could easily be overlooked in toxicology domain however other issues such as completeness and believability seem relevant to any domain.

2. Fusionplex [7] is a system that integrates information from multiple sources and also resolves data inconsistencies by use of fusion methods. For this specification a feature weight is identified and added to the database related to that source. Examples of features include: timestamp and cost. According to this method, inconsistencies are schematic differences between databases. For example, in one database we might have „salary" as an attribute and in another „income", where both still represent the same thing. Fusionplex uses criteria that are concentrated on information processing aspects rather than the data itself. Use of feature weight is also applicable

to toxicology domain. However, a main problem in predictive toxicology regarding the values conflict because of different originating sources remains untouched.

3. COLUMBA [8] performs the quality check by data cleansing. Errors in databases are considered of syntactic or semantic nature. Syntax errors are mainly domain or format violations in data entries and misspellings. Individual parsers perform syntactic cleansing such as dictionary lookup. Semantic errors affect the quality of the data significantly. These are resolved by using redundant information, which is possible in cases where another version of the same data source is available.

The limitation of the system COLUMBA lays on relying on the redundant data. This approach can not be extended entirely to Predictive Toxicology since there might be cases of sparse tables of data where there are no duplicated instances for model development in the initial data collection, so the quality assessment and validation process cannot be achieved. Another issue is how this redundant data can be qualified? What are the sources of this data? Some steps proposed by this method might be considered relevant, such as overcoming syntax and semantic errors, which are important issues in any database management systems.

4. The Information Quality Assessment Methodology [9] introduced by Richard Y. Wang, contains three components: product-service-performance, information quality assessment and IQ benchmark gap analysis. Each component contains further criteria in order to identify the best practice of company in production and delivery. The system proposes a method to measure information quality according to specific dimensions and is based on questionnaire. This approach is mostly based on data processing and it is user oriented. The evaluated outcome will differ between users.

Some criteria (free-of-error) need a rigorous metrics definition if applied in predictive toxicology. Such necessary metrics can explain for instance what sort of information is believable which still shows dependency. The other two components are entirely accessing the organizational performance in the sense of improving their products and services based on feedback from consumers.

5. The methodology for establishing and maintaining quality in data context [10] proposes five levels: test of completeness and emptiness; ranges and distributions; derived relationships; meaning and interpretation and hypothesis and discovery.

This approach could be a first step toward a reliable database management system in predictive toxicology. It is organization oriented although can be used as a stepping stone by any database administrator.

6. Data quality in predictive toxicology – identification of chemical structures and calculation of chemical properties by Helma [11] highlights some of the data inefficiencies and errors in toxicology databases and also draw some rules from a case study which was carried out in order to emphasize how some of the elements in experimental works could go under quality assurance. An example in toxicology databases relates to chemical compounds, which instead of chemical structure are identified by CAS Registry number and because of formatting or typing errors sometimes the compound cannot be identified. This approach emphasizes the idea of data representation rules in any source of toxicity database. Some of these rules are drawn from standard agencies for collecting, storing and processing such a data.

3. Data cleaning methods

Data cleaning techniques and procedures for noise removal could also enhance data quality. These methods could be applied to the data at different stages to address variety of data quality problems. At data collection stage, this could be in the form of removing duplicated records, missing values, spelling errors and outdated codes [12]. These techniques could also be used at data analysis stage. The purpose of data cleaning at this stage is to remove data errors in order to increase the quality for better classification models produced by machine learning and data mining algorithms. These techniques are based on outlier detection. Examples of some of these techniques are: cluster based, distance based and density based outlier detection.

In our experimental work data cleaning has been proposed at data collection stage. Also some metric has been proposed to detect outliers and suspicious values to increase the quality of the data for further modeling. These have been discussed further in the paper

4. Materials and methods

The contribution of our investigation at this stage is to highlight some common problems of data quality in toxicity prediction. Our current objective is the study of inconsistencies in data values and their affect on downstream QSAR modeling. We also rely on the data made available by research group of experts and the rules of compromise are already agreed on.

4.1. Case studies

Given the current facilities available for complex calculations, it seems that high confidence is implicitly awarded to data downloaded from online resources. The same applies to data generated by specialist software. We used the opportunity to study the DEMETRA data sets on some issues on data quality for large databases. We started with identification of descriptors sharing the same name and duplicated as generated by various software used by research laboratories involved in the project. We addressed the differences in data source values and also differences in performance of models developed from the same data sources. Data on five toxicity endpoints are provided by the DEMETRA project [13] for four different species: Bee, Daphnia, Trout, OralQuail and DietryQuail. For each dataset, values for six compound descriptors calculated by two specialist programs: ACD [14] and Pallas [15], have been considered. These programs calculate pKa, logP, logD values and also metabolites based on structural formulae of compounds. In the field of industrial pharmacy perhaps the most important physicochemical characteristics of compounds are their acidity or basicity (expressed by their pKa value), hydrophobicity and its dependence on pH (expressed by their logP and logD, respectively) [15]. Calculating accurate values of pKa, logP, logD and other chemical descriptors requires a great deal of work and use of specialized software.

For this work the number of chemical compounds present in each data set varies from 105 for Bee endpoint to 252 for Trout. Our aim was to highlight the variation of values for each descriptor produced from one program to another and also to compare any further quantitative differences between specific descriptors calculated by one

program with the value for the same descriptor and chemical compound generated by the other one. Then we compare the accuracy of basic classification models (developed using Weka [16]) built using input data presented for each endpoint by descriptors calculated by ACD and Pallas. Ten tables were investigated, two for each endpoint.

The aim of this experiment was to identify how the predictive models' quality is affected by hidden parameters such as source of data, subjective input characterization in running feature extraction algorithms *etc.* Comparisons of models performance will address variations, contradictions, reliability and deficiency issues.

4.2. Data pre-processing

For each dataset the same number of compounds has been selected. Data cleaning has been performed in the form of eliminating rows with missing values. The values for each descriptor in each row have been looked at in order to highlight deficiencies and wrong values. CAS number and chemical name for each corresponding ID number has been compared in both data sets to assure accuracy and homogeneity of data storage. Six common descriptors have been selected from both datasets. These are as follows: LogP, LogDpH3, LogDpH5, LogDpH7, LogDpH7.4 and LogDpH9.

The data have been divided into training set and testing set based on predefined rules (85% training, 15% testing) by DEMETRA project. Weka data mining tool has been used to develop models. The conditions of experiments for each endpoint containing two datasets were identical in order to assure an accurate comparison. Identifying aspects of data manipulation in further model development within our work, we aim to provide a clearer picture for scientists to perform quality assurance.

4.3. Comparison of global parameters and source value difference

For this experiment, after preparation, data files in their original format (Excel) were studied. We calculated the Mean, Min and Max for each input (LogP, LogDpH3, LogDpH5, LogDpH7, LogDpH7.4, LogDpH9) from both source files. At the second stage we compared each descriptors value for these parameters with its corresponding value from the other file (ex: parameter values for LogP from ACD were compared with same parameters from Pallas).

Then every row's value presented by ACD was subtracted from value presented by Pallas for the same chemical compound and same descriptor in order to measure their difference, as generated by different source program. Finally Mean, Min and Max were calculated for value differences as well. The results are shown in Table 1.

4.4. Comparison of models

Original data sets (prepared for training and testing) were used to develop Weka models based on the following algorithms: ClassificationViaRegression, BayesNet, MultilayerPerceptron, IBK, ZeroR, LMT, J48 and JRip. For performance of models study, two case studies have been considered. Firstly models obtained from training data (separated inputs from ACD and Pallas for same endpoint) were tested against test data sets. Secondly 10-fold Cross Validation has been used on training data. The accuracy of each model (one from modeling against testing set and one from modeling with Cross Validation) was recorded to identify which model suits which endpoint.

Table 1. Results of descriptors comparison by source for three endpoints

Pallas (Daphnia)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	−2.70	−6.54063	−6.54052	−7.85046	−7.89228	−9.10505
Max	11.7	11.6915	11.6915	11.6915	11.6915	11.6915
Min Value Difference	−7.4	−8.11613	−8.11602	−8.50406	−8.10928	−7.40214
Max Value Difference	2.21	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	346	51	51	417	417	143
ID of Max	418	418	418	418	418	418

ACD (Daphnia)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	−2.35	−5.4966	−5.8715	−6.499	−6.6644	−6.8685
Max	13.6	13.676	13.676	13.676	13.676	13.676
Min Value Difference	−7.40	−8.11613	−8.11602	−8.50406	−8.10928	−7.40214
Max Value Difference	2.21	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	143	143	143	143	143	143
ID of Max	90	90	90	90	90	90

Pallas (Bee)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	−0.95	−3.78509	−4.75384	−5.6052	−5.9931	−7.5105
Max	8.16	8.16996	8.16996	8.16996	8.16996	8.16996
Min Value Difference	−3.0	−3.00158	−3.00158	−3.00158	−3.00158	−3.27919
Max Value Difference	2.21	2.22028	2.21881	3.58602	3.80667	3.81636
ID of Min	192	382	457	373	373	373
ID of Max	146	146	146	146	146	146

ACD (Bee)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	−1.42	−3.4969	−4.2068	−5.0751	−5.2504	−5.8599
Max	8.26	8.1404	8.1404	8.1404	8.1404	8.1678
Min Value Difference	−3.0	−3.00158	−3.00158	−3.00158	−3.00158	−3.27919
Max Value Difference	2.2	2.22028	2.21881	3.58602	3.80667	3.81636
ID of Min	373	382	382	373	373	373
ID of Max	146	248	248	248	248	146

Pallas (Trout)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	−2.7	−6.54063	−6.54052	−6.5299	−6.51344	−9.10505
Max	8.68	8.68196	8.68196	8.68196	8.68196	8.68196
Min Value Difference	−7.4	−8.11613	−8.11602	−8.1055	−8.08934	−7.40214
Max Value Difference	2.69	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	346	51	51	51	51	143
ID of Max	93	93	93	93	93	93

ACD (Trout)	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
Min	−2.35	−5.4966	−5.8715	−6.499	−6.6644	−6.8685
Max	13.6	13.676	13.676	13.676	13.676	13.676
Min Value Difference	−7.40	−8.11613	−8.11602	−8.1055	−8.08934	−7.40214
Max Value Difference	2.69	4.00517	3.19537	3.58602	3.80667	3.81636
ID of Min	143	143	143	143	143	143
ID of Max	90	90	90	90	90	90

Other parameters from modeling can also be recorded. Table 2 shows classification accuracy for models obtained as described above, once using training set against test set and once using 10-fold Cross Validation with 8 algorithms on three endpoints.

Table 2. The algorithm classification accuracy for three endpoints

Endpoints	Algorithm accuracy against test set							
	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Trout(Pallas)	56.52	54.35	36.96	56.52	54.35	47.83	43.48	52.17
Trout(ACD)	63.04	65.22	47.83	63.04	56.52	58.70	43.48	60.87
Daphnia(Pallas)	42.50	47.50	40.00	40.00	42.50	45.00	40.00	42.50
Daphnia(ACD)	47.50	50.00	35.00	45.00	40.00	37.50	40.00	45.00
Bee(Pallas)	31.25	37.50	18.75	37.50	37.50	37.50	31.25	37.50
Bee(ACD)	31.25	37.50	31.25	31.25	25.00	31.25	31.25	37.50

Endpoints	Algorithm accuracy tested by 10-fold Cross Validation							
	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Trout(Pallas)	57.41	50.46	43.06	54.63	52.31	50.93	44.91	50.00
Trout(ACD)	56.48	49.07	51.39	54.63	51.39	56.02	44.91	51.85
Daphnia(Pallas)	44.12	47.55	38.24	43.14	48.53	43.14	44.61	45.59
Daphnia(ACD)	42.65	47.06	42.16	50.49	48.04	48.53	44.61	46.08
Bee(Pallas)	41.77	34.18	32.91	40.51	31.65	41.77	41.77	36.71
Bee(ACD)	35.44	36.71	39.24	39.24	35.44	44.30	41.77	39.24

4.5. Descriptor swap

Another experiment that has been performed on value comparison was to see the effect of one descriptor value (LogP) in the set of descriptor values in another file on the performance of the model. The logP descriptor of the ACD dataset for all the endpoints was swapped with the corresponding descriptor in Pallas and vice versa in both training and testing data set. Then with Weka and use of the same algorithms as for the previous experiment the models were trained once before swap and once after. As the result show in Table 3, the classification accuracy increases after LogP swap. Since the accuracy improves there should be better correlation between this individual descriptor value and the rest of the descriptors in each data set. This

Table 3. The Algorithm classification accuracy by 10-Fold Cross Validation after LogP swap

Endpoint	BN	MLP	IBK	CVR	J48	JRip	ZeroR	LMT
Trout(Pallas)-Before	44.12	47.55	38.24	43.14	48.53	43.14	44.61	45.59
Trout(Pallas)-After	57.41	52.31	44.44	54.63	50.93	50.93	44.91	49.54
Trout(ACD)-Before	42.65	47.06	42.16	50.49	48.04	48.53	44.61	46.08
Trout(ACD)-After	56.02	53.70	47.69	55.09	54.63	55.09	44.91	53.24
Bee(Pallas)-Before	41.77	34.18	32.91	40.51	31.65	41.77	41.77	36.71
Bee(Pallas)-After	41.77	31.65	31.65	41.77	30.38	40.51	41.77	40.51
Bee(ACD)-Before	35.44	36.71	39.24	39.24	35.44	44.30	41.77	39.24
Bee(ACD)-After	35.44	21.52	35.44	39.24	31.65	44.30	41.77	40.51
Daphnia(Pallas)-Before	44.12	47.55	38.24	43.14	48.53	43.14	44.61	45.59
Daphnia(Pallas)-After	46.08	47.06	38.73	44.61	44.12	46.08	44.61	45.59
Daphnia(ACD)-Before	42.65	47.06	42.16	50.49	48.04	48.53	44.61	46.08
Daphnia(ACD)-After	43.14	44.12	39.71	44.12	45.10	47.06	44.61	46.08

issue could also be considered as data pre-processing stage and also provide better understanding when it come to identify value variation bias for descriptors.

4.6. Results and discussions

We found the following deficiencies in data files:

Check of Input Values: there were number of rows in which the values for all columns (descriptors) were identical for specific chemical compounds. This might have happened as a result of a mistake in value generation by the software used due to the complexity of the calculation of the chemical compounds properties (ex: Trout data set). These values might be the default values for descriptors, which are generated when the exact measures for compounds attributes cannot be produced. For whatever reasons these values appear in the dataset, they need further consideration and study and they cannot be relied on.

We also found a contradiction between ID number and matching chemical specified by one program to another in the sense that the ID for the specific chemical was the same in both files but the matching name and CAS number were different. For example for endpoint Bee LD_{50}, in the file with ACD descriptors, chemical compound with ID = 450 = Allethrin has been given CAS no: 584-79-2 but in the file produced by Pallas, ID = 450 = 28 434-00-6 = s-bioallethrin, which in Toxnet comes with a different name for the same chemical having the same CAS: 284-792.

Moreover, a breach of the homogeneity rules was found: in the dataset for endpoint Trout, legend (descriptors definition) for Pallas is different from the other endpoints although for this work the descriptors were selected accordingly (ex: Pallas04 = LogDpH7 but for other endpoints Pallas05 = LogDpH7).

Also the number of significant places that represent values in each column and for every row is different, which shows inconsistencies of data representation. Table 1 presents the results of calculation for Min, Max values and their difference of the same descriptor for the same compound available in two data files related to the software used to calculate chemical descriptors and also shows the ID number of the chemical compound with the Min or Max value for the specific descriptor. What we found are significant differences between calculated values for the same descriptor presented by ACD and Pallas. In some cases, for example for endpoint Trout LC_{50}, the maximum values for LogP are 8.6 (Pallas) and 13.6 (ACD) and for OralQuail LD_{50} are and 8.1 (Pallas) and 13.6 (ACD) (see Table 1). This is almost double from one to another and flags out a significant warning, since information provided for this descriptor identifies compound solubility in water and ability to cross cell membranes and is therefore of high importance for toxicity prediction models.

Model Performances: descriptor value differences also create doubts of reliability. This problem applies to all descriptors and for all endpoints in DEMETRA datasets. In Table 2, the accuracy of models using various algorithms for classification is compared: values for the first experiment, which was model development based on training set using eight algorithms (see above) and validation against original testing set. The performance in general presents better results for data values generated by program ACD. Mean Square Error and Root Mean Absolute Error have been used to measure the errors of classification accuracy (not displayed here) are lower for models related to ACD data. This shows better correlations between ACD descriptors

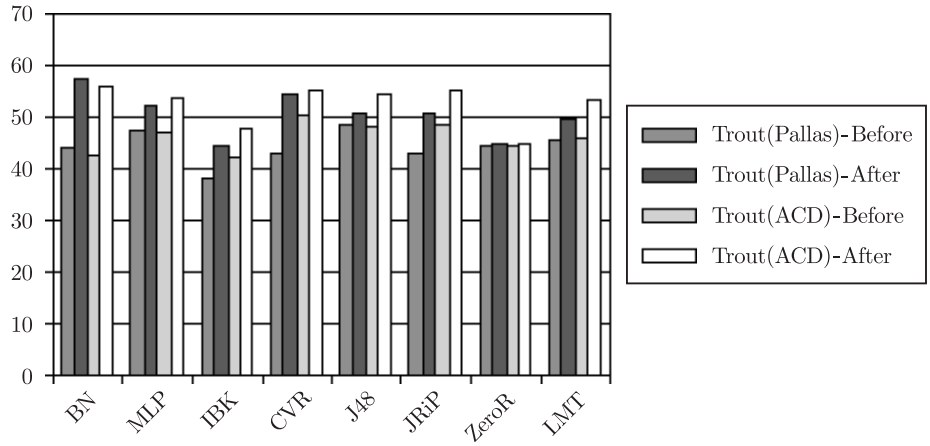

Figure 1. Comparison of classification accuracy of models after LogP swap for Trout endpoint

values and the toxicity output. Performance of the models has also improved with swapping descriptor between two dataset from ACD and Pallas (Table 3).

Figure 1 also shows the results of the LogP swap for Trout endpoint graphically. The first and third bar represents the value for classification accuracy before swapping LogP descriptors between two files (ACD, Pallas) and second and fourth bar shows the values after the swap. As the graph shows the difference between bars are very considerable especially for BN. CVR and JRip algorithms.

Range Margins' IDs: ID numbers for chemical compounds defining Min and Max value for same descriptor were also considered. If a chemical compound with specific ID number has the Min value for a specific descriptor in one data file, the same chemical compound should possess the same parameter property for all source files. For instance (Table 1) for endpoint Trout, the ID of compound, which has the minimum value for LogP (Pallas) is 346 but generated by ACD is 143.

Min-Max value difference between two columns (value for the same descriptor, one generated by ACD and one by Pallas) in the same row considerably vary (ex: for Trout LC_{50} endpoint vary by up to 8.1 unsigned numerical value).

5. Proposed criteria for data quality in predictive toxicology

Figure 2 shows values variation for LogP between the two programs (data for OralQuail LD_{50} endpoint). There are number of big peaks in the graph for values calculated by both programs which clearly identify the presence of outliers. As it shown the values follow same pattern but in different proportion. This again depends on the computer program calculation default values setup, which is not the same in two programs. From this experiment we propose as property for data quality the definition domain for each variable (a value range for each descriptor) and decide that we just accept the values in this range and categorized the peaks outside the range as outliers so they could be studied separately. This bias could be proposed as metric for every descriptor considering the measurements of every descriptors confidence interval for each endpoint and acceptance of the values within this range. Later we need to

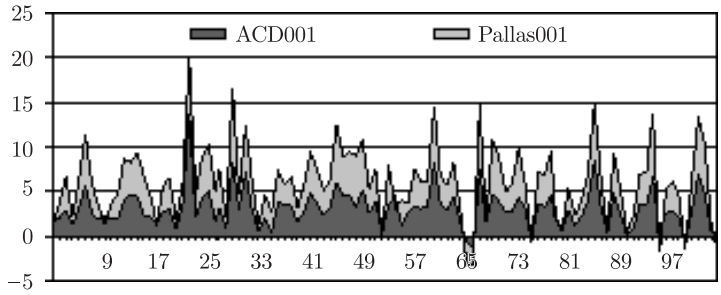

Figure 2. Comparison of LogP variation values presented by ACD and Pallas for OralQuail

Table 4. Calculated variance for OralQuail

OralQuail	LogP	LogDpH3	LogDpH5	LogDpH7	LogDpH7.4	LogDpH9
VARP (ACD)	4.83210	5.52960	5.62856	6.18880	6.26632	6.23918
VARP(Pallas)	4.87138	7.21954	7.29682	7.51035	7.51433	7.37912

define a method to describe how the outliers could be modeled separately and how we can combine these models with the results of the training the rest of the data.

Table 4 shows the variance VARP calculated for descriptor values obtained by using ACD and Pallas for OralQuail LD_{50} endpoint according to formula: $\frac{\sum(x-\overline{x})^2}{n}$ where x is a sample Mean and n is a sample size. The variance values are greater for values produced by Pallas, which shows bigger distribution with a negative impact on the model development. The descriptor variance qualifies as a meaningful property of the source values. Noise in data identified by rows with the same values in each column could be another measurement for signaling wrong data inputs. These rows should be recalculated.

A correlation of the margins (Min and Max values) for each descriptor as calculated by different software represents a quality flag variable as well. If these extreme values (generated by various sources) for each endpoint do not belong to the same compound, then that particular descriptor needs further study. This is especially requires further consideration for descriptors (*i.e.* LogP) that are likely to be included as inputs for models based on feature extraction algorithms.

Descriptor swap (LogP) increased the classification accuracy. This showed the change of input balanced the model, which also can be used in defining bias for descriptors min and max values.

5.1. Quality processing flow chart for proposed metrics

Based on empirical results obtained from studying the five toxicity datasets, we propose a data quality assessment process. Figure 3 shows this necessary process to prepare data for further modeling based on highlighted defects in our experimental work at this stage. Note that investigation was carried out on internal data and the proposed process has been based on discovered results. A final version of the procedure is still under development.

As the figure shows, first input values from data sources are checked one by one. At the second stage based on quality check (Q_1) rows with missing values are identified and eliminated. Then the rows with disguised missing values are flagged

(Q_2) (these are the rows with same value in all the columns). The values in each column for every chemical compound are then checked for out of range values based on (Q_3). Then comparison of minimum and maximum values are performed (Q_4). At this stage if the data is rejected the process ends otherwise the data is modeled. After first modeling the value difference between descriptors are calculated. If the value difference is considerable, LogP between two dataset would be swapped and data would be trained for the last time.

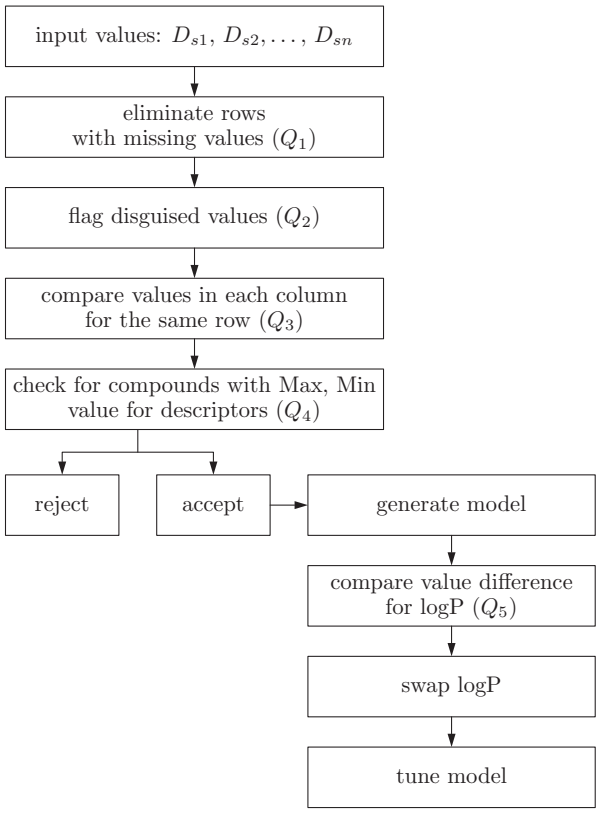

Figure 3. Data quality assessment procedure

5.2. A quality assessment algorithm for data quality procedure

In Figure 4 we propose a quality check and assessment algorithm for the above procedure. The proposed algorithm could be improved and extended to provide further quality checks. At this stage the main aim was to direct our attention to the first stage, error identification defects and highlighting defects and propose possible ways of discovering and overcoming these in toxicology data.

6. Conclusions and future work

Considering data quality parameters and criteria identified by our study and the experimental work presented above, some issues related to data quality have been highlighted, which indicate the need for a framework for quality assessment

Double data source
Input: D_s: Data Source, R_o: Result Output (data processed, ready for modeling), R_w: Instance(compound, row), D_c: Descriptor(column), result: R_s (final model), Quality Metrics: Q_1=missing values in rows, Q_2=column values are same in one row, Q_3=values for the descriptor in each row is out of range Minv→Maxv, Q_4=flag if Min and Max value for same descriptor in two files do not belong to the same compound, Q_5=bias for value difference between same descriptor value for same compound in D_{s_1}, D_{s_2}, D_{s_n}... In our example data sources are ACD and Pallas.

Clean the data
//check for rows with missing values and eliminate
//check for rows with same value in each column and eliminate
//compare if value for each descriptor (column) and every row falls within bias (value range)
Start: SearchSheet (ACD & Pallas)
Foreach (Result as sheet→R_o)
For (i=0; i<count (R_w); i++)
If (R_w=(Q_1))
Delete R_w else
For (j=0; j<count (D_c); j++)
If (D_c=(Q_2))
Flag (error): "disguised data" else
If not R_w=(Q_3) & D_c=(Q_3)
Flag (error): "suspicious values" else
If not R_w=(Q_4) & D_c=(Q_4)
Flag (error): "Min, Max do not belong to same compound" else
Display R_o
End

Generate model
//check similar fields; if value difference for same descriptor (logP) and for same compound is high, then train model, produce result, swap logP, train again.
Input: R_o+added new column which shows the difference between two values (logPPallas-LogPACD=D_{s_w}), LogP descriptor=DLogP
//generate model with cleaned data
Start: generate model (ACD, Pallas)
//swap logP and generate again
Foreach (Result as sheet→R_{s_w})
For (j=0; j<count (D_c); j++)
If (R_{s_w}=Q_5)
Swap (DLogP)
Display R_{s_w}
Generate model
End

Figure 4. Data quality assessment algorithm

and measurements. The experimental work has identified some deficiencies related to data values and presentation. All highlighted data defects have direct effect on QSAR model performances. Further work would involve investigation into datasets, values for chemical compounds descriptors and relationships between attributes.

Acknowledgements

This work is partially supported by the EU EP5 project DEMETRA ([13]). DN acknowledges the support of the EPSRC project GR/T02508/01. LM acknowledges the support of the CSL sponsorship.

References

[1] Eriksson L, Jaworska J, Worth A, Cronin M, McDowell R and Gramatica P 2003 *Environ. Health Persp.* **111** (10) 1361

[2] *http://toxnet.nlm.nih.gov/*

[3] *http://www.ncbi.nlm.nih.gov*

[4] Warr W A 2003 *IUPAC Project Meeting: Extensible Markup Language (XML) Data Dictionaries and Chemical Identifier*, NIST, USA

[5] Hunter A 2006 *Data & Knowledge Engineering*, Elsevier, **57** 221

[6] Naumann F and Roker C 2000 *Proc. Int. Conf. on Information Quality (IQ2000)*, Cambridge, Mass, pp. 148–162

[7] Anokhin P and Motro A 2003 *Technical Report ISE-TR-03-06*, Information and Software Engineering Dept., George Mason University

[8] Rother K, Muller H, Trissl S, Koch I, Steinke T, Preissner R, Frommel C and Leser U 2004 *COLUMBA: Multidimensional Data Integration of Protein Annotations*, Germany

[9] Yang L, Strong D and Wang R 2002 *Information and Management* **40** (2) 133

[10] Tap R 1999 *Conf. on Information Quality*, MIT Sloan School of Management, pp. 209–219

[11] Helma C, Kramer S, Pfahringer B and Gottmannn E 2000 *Environ. Health Persp.s* **108** (11) 1029

[12] Xiong H, Pandey G, Steinbach M and Kumar V 2006 *IEEE Trans. on Knowledge and Data Engineering* **18** (3) 304

[13] *http://www.demetra-tox.net*

[14] *http://www.acdlabs.com*

[15] *http://www.osc.edu/ccl/pallas.html*

[16] *http://www.cs.waikato.ac.nz/ml/weka*