

Dataset quantity reduction with RDBMS built-in mechanisms as specific data preprocessing for data mining algorithms

Magdalena A. Tkacz

Institute of Computer Science, University of Silesia, Sosnowiec, Poland

Abstract

Data mining algorithms are rather time and resources consuming during it's work. That means, that the only way to speed up computations and to receive appropriate results is suitable dataset preparation before processing. That, inter alia means, that dataset processed with data mining algorithms should be informative and should be properly prepared for automated processing (for example dataset should not contain outliers). In this paper a method for filtering data with RDBMS build-in mechanisms for further – data mining processing, is shown.

Keywords: RDBMS, data preprocessing, data mining, gene expression, Affymetrix, genechip

1 Background and motivation

Today there is no problem to store, index and search for known data or necessary information. Technical possibilities reached such a level of development, that there is no problem in storing, searching and filtering large amount of data. However, now the problem is to find out an information which should be found. Moreover – why did not search for different relationships, dependencies or similarities among data? Why do not have a chance to discover unexpected or surprising facts? For that purpose there is a relatively new field of researches in computer science – data mining (DM) and subsequently – knowledge discovery from data (KDD) (Hand 2005).

First of all, let's make some explanation and assumptions what does under that terms will be understood. Particular terms can be described as follows:

- *data mining (DM)* is referred to as analyzing large amounts of data and selecting an information that is relevant to considered problem and interesting from our point of view. This process can be conducted with a support of different methods, such as inferential statistics, descriptive statistics, soft computing and artificial intelligence methods. The difference between data mining and data analysis is a type of information presented to the user. In data analysis

we obtain some patterns, trends regarding specific user query (some projections, intersections and comparisons), in data mining user can obtain some dependencies and relationship not easily observed at first sight. Data mining gives more information than data analysis and is often a part of KDD process. With data mining we have a chance to discover interested, but sometimes unexpected relationships or data conformity.

- *knowledge discovery from data (KDD)* is referred to different automated methods which helps search large (or huge) volumes of data. The most important in this process is to find directly unknown information or unexpected patterns and relations among data, but indirectly – relationships among objects or processes of interests which are represented by the data. This should help us to understand and describe objects or processes which are related to data, because we are looking for some kind of knowledge concerning the analyzed data – hidden inside, and within data relationships.

The difference between data mining and knowledge discovery from data is rather in approach than in used methods: in data mining we analyze large amount of data looking for the interesting problem from our point of view. In knowledge discovery from data we are trying to find out something new – possibly yet unknown, maybe sometimes quite novel and unexpected. We can consider knowledge discovery from data as a method of finding dependencies, but ... rather as a result of data scanning and processing, than object or process observation. Unfortunately, both data mining and knowledge discovery from data are time and resources consuming, so very important for data mining process efficiency is to include in our searches only necessary, informative and appropriate data (Larose 2005; Larose 2006). All necessary data preprocessing are often referred to as ETL process (it is acronym from Extraction, Transformation, Load). To make possible automated data analysis (and data mining and knowledge discovery from data with its wide range of algorithms (statsoft)) we (inter alia) have to separate the outliers (Larose 2005). It is the name of certain data which significantly differ from each other, which do not fit to the rest of data. When we have a dataset “Age” we are expecting that all data should be – for example from the range of 0 to 130. It is clear for the human being, that a numbers like 2 321 or -23 are inappropriate in this context. But computer will not detect such a value as a “bad value”, so suitable initial dataset preparation is crucial for obtaining proper results.

2 Dataset preparation – basic information

Both in data mining and knowledge discovery from data we are screening large sets of data. It is rather obvious that we have to utilize some kind of data storage. In fact, it may be any existing database system. Unfortunately, having at least a dozen of Relational Database Management Systems (RDBMS), they are not a quite good source for data mining or knowledge discovery from data, because they are developed for transactional databases. They should maintain security for data storage, data consistency and assure efficiency for both writing and reading data. In fact, in data mining we do not have to think about writing and modifying

TABLE 1: Exemplary microarray reading (from resources of Microarray Lab in Chair and Division of Molecular Biology, Medical University of Silesia, Poland). Spec1 and Spec2 are control samples, Samp1 and Samp2 are examined samples.

ProbesetID	Spec1	Spec2	Samp1	Samp2
205476_at	6.625	5.394	10.503	10.386
202859_x_at	9.741	10.567	9.660	4.854

data – we analyze a certain data “snapshot”. So, we can rebuild data structure in that way that this structure will be most efficient for searching. One of the best suitable data structures for searching are trees. For that reason, for data mining a special data structures *star* or *snowflake* are created basing on tables in relational database management systems. Nowadays there are no rareness the databases which have more than terabytes in size. But in data mining it is not necessary to take into account all database data. Often we need only part of data – it is reasonably to analyze data which potentially have – maybe an intuitive relationship. As mentioned above – it is worth to “shot” and separate outliers for distinctive, not quite automated analysis.

So, under assumptions made above, we do not need to copy all data into new, to-be-mined structure. When we know something about data itself, and we understand for what we are looking for – we could select most interested, most informative data from given databases. Then, when we are be able to reduce mined data, we have a chance to mine all necessary data, and in the same time to have a chance not to worry about amount of data and efficiency of mining algorithms. In next section it will be shown how quantity of data can be reduced on the basis of some additional information about data and expected experiment results.

3 Dataset preparation – dataset quantity reduction without information loosing

Below will be shown how dataset quantity can be reduced before processing with data mining algorithms with structured query language (sql) basing on knowledge about experiment goal. The goal of experiment is to find out differentially expressed genes or transcripts which expression values differ from each other in whole Affymetrix microarray readings. Let’s assume that we have four microarray readings: two “specimens”, and two analyzed samples. Each Affymetrix microarray reading contains readings for 22 283 spots. In fact, some of them are “control” spots, so we can regard that we have readings about 22 000 transcripts (genes) for each microarray. Having four microarray readings – finally we have “at the beginning” an array (table of data) with dimension of 5 columns (transcript/gene ID and 4 microarray readings) and 22 283 rows (see table 1).

For further analysis and interpretation we need information, for what about every transcript (gene) is responsible for – we need to reach the Affymetrix database (Affymetrix) and we have another data table (see table 2), with 8 columns

and 22 283 rows, with a lot of text information.

Because there are a one-to-one relationship between transcript id (written as ProbeSetID in tables), these two tables (from microarray reading and those from Affymetrix database) from logical point of view can be treated as one big table with 22 283 rows and 12 columns. Because of the goal of the experiment – finding out differentially expressed genes – all discussions concern only expression (microarray) readings, not all information – the rest of information is not explained here in more details. This papers have limited size, so only fragments of the tables are shown. For the gene expression analysis we need to find out, for which transcript id's in the readings the expression values will differ from each other. As informative can be regarded only such set of expression reading (rows) which are almost equal for the specimens (Spec1 and Spec2), and almost equal for the samples (Samp1, Samp2), but at the same time – quite different for specimens and samples. There is a known and useful measure of dispersion – it is a variance, so we compute a variance for samples, specimens and for samples and specimens together and add them to our datatable (see table 3).

Now, we can use a simple SQL query to reduce data quantity for further analysis. First of all, taking into account that we have two samples and two specimens, we have to know from which readings an average can be computed (for which samples and specimens an average is a “proper” value – similar and representative to the rest values).

To check out “quality” of readings we are looking for those readings, for which variance value will be **smaller** than a certain, fixed value. Here, important thing is, that readings should be more or less identical, what means that variance of samples and variance of specimens should be a small number. It is necessary to be assure that computed, average value of two readings is representative. Let variance be equal to 0,15. For that threshold, after executing a query

```
SELECT Readings.VarSamp, Readings.VarSpec
      FROM Readings
     WHERE (((Readings.VarSamp)<0.15)
           AND ((Readings.VarSpec)<0.15))
```

remains 17 025 readings of transcript expression (rows). Result of that query has been stored in a new datatable `ReadingsLess0.15`. Next step, (remembering about type of data analysis) is to select all records that potentially have different values of transcript expression in all readings. To do so, it is enough to select all those records for which variance of all readings will be **greater** than some given value. A certain, fixed value will depend on life scientists experience and on desirable results and desirable precision. For that article's purpose a variance threshold has been set to 0,2. After executing that query

```
SELECT [ReadingsLess0.15].[VarSampSpec]
      FROM [ReadingsLess0.15]
     WHERE ((([ReadingsLess0.15].[VarSampSpec])>0.2))
```

remains 292 readings of transcript expressions for further, automated analysis.

Few examples about different final dataset quantity for data mining algorithms in dependency of variance values are shown in table 4.

TABLE 2: Records (two exemplary of 22 283) from Affymetrix [Affymetrix] used for data interpretation

ProbesetID	Gene Symbol	Gene Title	Pathway	Pathway Hyperlink	GO BioProc	GO MolFun	GO CellComp
205476.at	CCL20	chemokine (C-C motif) ligand 20	-	-	chemotaxis, inflammatory response, immune response, signal transduction, cell-cell signaling, defense response to bacterium	cytokine activity, protein binding, chemokine activity	extracellular region, extracellular space
202859_x.at	ATP5L	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit G	Electron Transport Chain	<i>see * below this table</i>	ATP biosynthetic process, transport, ion transport, ATP synthesis coupled proton transport, proton transport	protein binding, hydrogen ion transporter activity, metal ion binding, hydrogen ion transporting, ATP synthase complex, coupling factor rotational mechanism	mitochondrion, proton-transporting two-sector ATPase complex, proton-transporting ATP synthase complex, coupling factor F(0)

* Pathway hyperlink: http://www.genmapp.org/HTML/_MAPPs/Human/Hs_Contributed_20051123/metabolic_process-GenMAPP/Hs_Electron_Transport_Chain/Hs_Electron_Transport_Chain.htm

TABLE 3: Microarray readings with additional parameter – variance added. Spec1 and Spec2 are control samples, Samp1 and Samp2 are examined samples. VarSamp is a variance of Samp1 and Samp2 (two readings), VarSpec is a variance of Spec1 and Spec2 (two readings), VarSampSpec is a variance of Samp1, Samp2, Spec1, Spec2 (four readings).

ProbesetID	Samp1	Samp2	Spec1	Spec2	VarSamp	VarSpec	VarSampSpec
205476_at	6.625	5.394	10.503	10.386	0.758	0.007	6.812
202859_x_at	9.741	10.567	9.660	4.854	0.342	11.55	6.761

TABLE 4: Exemplary dataset quantity reduction (from 22 283) with different variance thresholds

Final dataset quantity	Variances for specimens and variance for samples (less than)	Variances for specimens and samples (more than)
102	0.05	0.2
40	0.05	0.3
196	0.1	0.2
79	0.1	0.3
292	0.15	0.2
154	0.2	0.3
110	0.2	0.35

Basing on examples presented above, it is clear, that such a simple operation as appropriate data selection (concerning information about experiment and data themselves) **before** further analysis can drastically reduce the quantity of data for further analysis. This can be done without losing interesting experiment information contained in data.

4 Conclusions

We do not have possibility to decrease algorithm complexity. For that reason in some cases, where alternative algorithms are not present, we have to make very resource and time consuming analysis. Although, even if we do not have another algorithm, we can significantly reduce the time we need for data analysis by a special data preparation carried out **before** advanced analysis. In this paper has been shown, that with some understanding of analyzed data we can significantly reduce the amount of data which have to be processed – without losing information. It has been also shown, that we are able to reduce the quantity of data – from almost 22,000 to about 300 – it is approximately about 100 times smaller number of data to process. One has to remember, that even when we are searching only microarray readings, some interested features or criterion for data mining is taken from the second table (see table 2), so subsequent filtering could effect further data reducing. In that way we could significantly reduce the amount of data to be mined – in fact, taking into account not only gene expression data, we have to search and compare not only numbers, but also a lot of texts which, generally is

more challenging task (Larose 2007).

However, there are still a few problems to solve in such approach:

- How to make possible a more automated threshold determination? (Maybe basing on some kind of variance distribution ?)
- How to measure a dispersion of values ?
- How to find out outliers?

The proposed method does not give us possibility to find out outliers (a single reading which is not similar to the rest). For more results (two or more readings) using variance as a dispersion measure, will not give us such possibility when – for example – a set of dozen values contains one, or two detachable values. When variance is applied, we will be not able to find such a value (values), even though we know, that such outstanding value (values) may have a significant effect on computed average value.

Acknowledgements

The research has been partially supported by the project “Decision support – new generation systems” of Innovative Economy Operational Programme 2007-2013 (Priority Axis 1. Research and development of new technologies) managed by Ministry of Regional Development of the Republic of Poland.

References

- [Affymetrix] www.affymetrix.com
- [Hand 2005] HAND D., MANNILA H., SMYTH P: Eksploracja danych. WNT, Warszawa 2005
- [Larose 2005] LAROSE D.: Discovering Knowledge in Data: An Introduction to Data Mining, Wiley 2005
- [Larose 2006] LAROSE D.: Data Mining Methods and Models, Wiley 2006
- [Larose 2007] LAROSE D.: Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage, Wiley 2007
- [statsoft] <http://www.statsoft.com/textbook/stathome.html?stdatmin.html&1>
- [sql] <http://www.w3schools.com/sql/default.asp>

