Author: Witsenburg, Tijn
Title: Hybrid similarities : a method to insert relational information into existing data mining tools
Date: 2012-12-20

Chapter 1

Introduction

1.1 Motivation

Little could the creators of the first computers suspect the impact they would have on society nowadays. Invented as machines to help men with difficult calculations, they soon evolved into convenient tools for storing and sharing huge quantities of information. Statisticians Lyman and Varian suggested in 2003 that worldwide data volumes will double every two to three years [60]. These data can be anything like the results of scientific research, internet, customer behaviour, video's, computer games, and so on. Besides the amount of data that is stored, also the amount of data that is consumed can give somewhat of a perspective to the extend to which information influences our behaviour. Bohn and Short stated that the amount of data consumed by Americans in 2008 was 3.6 zettabytes, which is $3.6 \cdot 10^{21}$ bytes, or $1,000,000,000,000$ gigabytes [6].

The collecting and storing of data has an increasing role in society where everyday activities are more and more being digitalised. Examples are: people using social network sites create a network of people, super markets that use 'customer cards' to monitor customer behaviour, public transport companies that use electronic checking in and out, and governments that stimulate people to fill in their tax form electronically.

All this sort of behaviour creates huge databases, containing potentially interesting information for sociologists, marketers, demographists, and various different types of scientists. While this enormous amount of data gives scientists many opportunities to conduct research, it is difficult for them to actually extract the useful information. Data nowadays can consist of billions of records, or observations, each having many different variables (also known as attributes). The large amount of attributes, or high-dimensionality of the data, makes it impractical to compare each pair of attributes. Also, the large amount of records limits the amount of times that an algorithm can pass through the whole data set. Classical statistical analysis methods are not capable of handling such amounts of data. Therefore a new type of methods needs to be invented.

The field of research that addresses this problem is known as Knowledge Discovery in Data (KDD). The most important step in the KDD process is known as data mining. Data mining is concerned with "applying computational techniques (i.e. data mining algorithms implemented as computer programs) to actually find patterns in the data", as defined by Džeroski [20]. These 'patterns' that Džeroski describes are not explicitly defined and depend on both the task that the data mining tool needs to perform, and the type of data it is performed on.

A data mining task can be seen as "what the user wants to know". For instance, if the user wants to predict a label for an element in the data set, then the data mining tool should be designed to do exactly that. Other examples of data mining tasks can be to divide the elements into groups of similar type, or to find combinations that occur often. In each case, of course, the pattern to be found is defined in a different way. On a more abstract level, these data mining tasks all search for the same: some underlying structure in the data that gives rise to patterns between elements.

The patterns to be found also depend on the type of data. Earlier data mining tools work only on data that is in a single table. Here, every row is considered one object of interest and every column an aspect, or attribute, of that object. Unfortunately, data do not always come in the same shape. Though there are many databases that consist of data in the form of a single table, many more do not. Examples of different types of data are relational data, graph data, video data, time sequence data, et cetera. Since data mining tasks in general only work on one type of data, for every other type, new data mining tools need to be invented.

To make things more complicated, a database could consist of a combination of more than one type of data. A very popular example nowadays are social networks, where all information about the people in it (e.g. their interests and likes) could be placed in a single table, but the friend structure is considered relational data. A second example would be a data set with scientific papers. Here the content (words that appear in the text) can be placed in a single table, and there is also relational information in the form of citations between papers, or papers having an author in common.

Data mining tools tend to use only one type of data. When such a data mining tool is applied to a database with multiple types of data, it cannot use all the information available in this database. In theory, this does not necessarily need to be a problem. The solutions that are found with data mining tools that use only part of the data can be just as correct, as solutions found with a data mining tool that uses all the information in the database. It could even be so that there is just as much confidence in the quality of the found solutions. Despite all that, in practice, a data mining tool that is able to use all the information in the database, could explore a much wider search space and would therefore, in potential, be a much more powerful data mining tool.

In this thesis, a data mining tool that uses more than one type of data is considered a hybrid data mining tool. When a hybrid data mining tool is created for a specific combination of data types, it will probably be very powerful on a database that has that combination of data types, but not perform well on another database that consists of a different combination of data types, should it even be able to use this other database as input.

There are many different types of data. The amount of different types of databases that can be created from combining multiple types of data, is exponential in relation to the amount of data types and thus very big. It is an incredibly elaborate task to create a new hybrid data mining tool for every possible combination of data types and every data mining task. So, in order to create hybrid data mining tools, a more general approach is needed. One such approach could be to create a method that incorporates one type of data in a data mining tool which is created for another type of data.

In this thesis, we propose a method that does this by inserting relational information in data mining tools that are created for single table data and use a similarity measure to compare the elements in the database with each other. It can also be used for other types of data, as long as the data mining tool uses a similarity measure. This leaves a wide range of possible data mining tools that can be enhanced by using this method. For various data mining tools, it has been implemented and tested. Although the results varied, overall, this method outperformed the original data mining tools.

1.2 Thesis Outline

This thesis consists of four parts: Foundations, Implementations, Analysis and Conclusions. Part I *Foundations* starts in Chapter 2 where the current state of data mining is described. It concentrates on the aspects that are important for this thesis. Then in Chapter 3 the method proposed in this thesis is explained. This is done by firstly describing the kind of data for which this method is created, and secondly, precisely describing the method itself. For the latter, two different approaches are used: one graph-based approach and one matrix-based approach. Also in this chapter, the method is compared to other methods that can combine different types of information. The final chapter in this part, Chapter 4, gives an intuitive motivation on why the proposed method could work better than already existing data mining tools and from this extracts some characteristics of the data for which this method could be useful.

The method proposed in this thesis can be used to improve already existing data mining tools. Part II *Implementations* describes the data mining tools for which this is done and gives the results on real data for the improved methods in comparison to the original data mining tools. Chapter 5 does this for agglomerative hierarchical clustering, Chapter 6 does this for k-means and Chapter 7 does this for KNN-classification.

Part III *Analysis* takes a closer look at the exact working of the method. First, in Chapter 8 the mathematical benefits of the proposed method are analyzed. Then, in Chapter 9 alternative possibilities for the proposed method are constructed and compared them with each other.

Finally, Part IV *Conclusions* gives an overview of all the conclusions that can be drawn from this thesis.

1.3 Publications

Several of the chapters in this thesis are based on publications. Listed below are the publications for each chapter.

Chapter 4: Motivation

This chapter is based on a paper presented at MPS'10:

> Hendrik Blockeel and Tijn Witsenburg. Exploiting homophily in unsupervised learning. In *Workshop on Mining Patterns and Subgroups*, Leiden, The Netherlands, 2010.

Chapter 5: Agglomerative Hierarchical Clustering

This chapter is based on a paper presented at MLG'08:

> Tijn Witsenburg and Hendrik Blockeel. A method to extend existing document clustering procedures in order to include relational information. In S. Kaski, S. Vishwanathan, and S. Wrobel, editors, *Proceedings of the 6th International Workshop on Mining and Learning with Graphs*, Helsinki, Finland, 2008.

Chapter 6: K-means and K-medoids

This chapter is based on a paper presented at ISMIS'11:

> Tijn Witsenburg and Hendrik Blockeel. K-means based approaches to clustering nodes in annotated graphs. In M. Kryszkiewicz, H. Rybinski, A. Skowron, and Z.W. Ras, editors, *Foundations of Intelligent Systems, ISMIS'11 Proceedings*, pages 346–357, Warsaw, Poland, 2011.

Chapter 8: Reducing the Impact of Content Variability

This chapter is based on a paper presented at the 19th International Symposium on Methodologies for Intelligent Systems:

Tijn Witsenburg and Hendrik Blockeel. Improving the accuracy of similarity measures by using link information. In M. Kryszkiewicz, H. Rybinski, A. Skowron, and Z.W. Ras, editors, *Foundations of Intelligent Systems, ISMIS'11 Proceedings*, pages 501–512, Warsaw, Poland, 2011.

Other Publications

The following publications on related subjects were co-authored during the PhD thesis:

Hendrik Blockeel, Tijn Witsenburg and Joost N. Kok. Graphs, Hypergraphs and Inductive Logic Programming. In P. Frasconi, K. Kersting and K. Tsuda, editors, *Proceedings of the 5th International Workshop on Mining and Learning with Graphs*, pages 93–96, Firenze, Italy, 2007.

Hendrik Blockeel, Hossein Rahmani and Tijn Witsenburg. On the importance of similarity measures for planning to learn. *Proceedings of the 3rd International Workshop on Planning to Learn*, Lisbon, Portugal, 2010.