The Knowledge Engineering Review, Vol. 17:2, 197–203. © 2002, Cambridge University Press
DOI: 10.1017/S026988890200022X Printed in the United Kingdom

5th European conference on principles of knowledge discovery in databases

FRANS COENEN

Department of Computer Science, The University of Liverpool, Liverpool L69 3BX; e-mail: frans@csc.liv.ac.uk

1 Opening

PKDD 2001, the 5th European Conference on Principles of Knowledge Discovery in Databases (PKDD), was held in Freiburg, Baden-Württemberg, Germany, this year (Monday 3 to Thursday 7 September), and co-located with the 12th European Conference on Machine Learning (ECML 2001). The proceedings comprised two volumes, one for PKDD (De Raedt & Siebes, 2001) and one for ECML (De Raedt & Flach, 2001); and form part of the Springer Lecture Notes on Artificial Intelligence (LNAI) series. The conference was held in the University buildings in the centre of the old town. Freiburg and the surrounding area were for many years part of the Austro-Hungarian empire and thus the university was described to us as being one of the oldest Austrian Universities.

The conference was opened by Luc De Raedt of Albert-Ludwigs University, and Arno Siebes of Utrecht University (PKDD programme co-chairs). For those who are interested in submission statistics, 78 papers were submitted to PKDD, 117 to ECML and 45 as "joint papers". Out of the 240 submitted papers 94 were accepted. Over half (54) of the accepted papers were initially given only a conditional acceptance, full acceptance being gained only after revisions recommended by the reviewers had been completed. I thought this system of conditional acceptance, despite the additional reviewing that this entailed, was a good one; other delegates were not so enthusiastic (but then our own paper was given a conditional acceptance). Glancing round the room I estimated that some 300 delegates were present.

The opening address was followed by a presentation by Stefan Wrobel of the Otto Von Guericke University of Magdeburg; the first of a number of invited talks. The subject of Stefan's talk was "Scalability, search and sampling". One of the principal problems that the KDD community is addressing is that identifying patterns in small data sets is trivial, however when the data sets become very large, wide and/or structurally complex specialist techniques are required. One solution is to concentrate effort on a part of the data set (a "sample"). However, we then need to know whether the sample is truly representative or not. Stefan gave an overview of the difficulties associated with scalability and by extension sampling. He went on to review the effectiveness of sampling and to consider how we might give some probabilistic guarantees concerning the effectiveness of our sampling strategies.

2 Sessions 1 and 2

Following Stefan's presentation we stopped for coffee after which the conference continued with two parallel sessions and a workshop and tutorial programme. There were two tutorials on the first day: "Development and application of Ontologies", "Bayesian intelligent decision support"; and two workshops "Semantic web mining" and "Machine learning as experimental philosophy of science". I thought the last workshop sounded particularly interesting (and a "bit different"), but decided to attend the conference session on clustering (the parallel session was on machine learning and induction).

Clustering is an important concept in KDD. The idea being to identify groupings/neighbourhoods of data in some N-dimensional "data space". The best known clustering algorithm is founded on the "k-nearest"-neigbour-technique, however there is no consensus on the "optimal" clustering algorithm.

The difficulty with clustering is how best to divide up the data space to give the "best" groupings for a given domain/application.

The session on clustering comprised four papers. The first paper was presented by Maria Halkidi from the University of Athens. Maria commenced by giving a review of the cluster challenge – finding the "real partitions" – and emphasised that visualisation of the end result is important. She then proposed a "validity-indexing algorithm" with which to evaluate existing clustering algorithms. Applying this algorithm to K-means, DBScan and CURE indicated that the first two performed better than the last.

The second paper in the clustering session was given by Chien-Yu Chen of the National Taiwan University. Chien-Yu presented a clustering algorithm founded on Wright's 1977 Gravity Theory equation. For readers who have not heard of this theory it does not mean that all data falls to the "floor" of the data space! Instead individual data items are considered to behave as celestial bodies, each with some notion of mass and consequently gravitational "pull". Chien-Yu suggested that altering the distance term in Wright's equation meant that some undesirable clusterings could be avoided. The results presented indicated that the algorithm worked well.

Véronique Ventos of the Paris-Sud University described a hierarchical classification algorithm. Classification is a form of clustering influenced by the class hierarchy concept found in object-oriented analysis, and programming languages such as Smalltalk and more recently Java. Véronique described a two-stage approach: (1) commencing with "semi-structured" data a coarse classification is derived, (2) a richer classification is then derived by focusing in on part of this classification (so that overall complexity remains under control). As I understood it each of these stages has its own dedicated language (L1 and L2), and in her conclusions Véronique hinted at a third language (L3). I was interested to see that one of Véronique's co-authors was Marie-Christine Rousset who was noted for her work in the late 1980s on verification and validation (V&V) of rule bases and the definition of inconsistency.

The final presentation in this session, on "conceptual clustering functions", was presented by Céline Robardet from the University of Lyon. Conceptual clustering is concerned with the integration of clustering and the interpretation of the results. Céline described techniques (functions) for evaluating the proximity between objects and the quality of partitions and presented some empirical results which demonstrated the significance of the approach.

3 Sessions 3 and 4

After lunch I attended the session on "Similarity and distance", again a clustering-related topic in that data that are in some sense similar should be grouped together in the same way that data that are in some sense distant from each other should not be grouped together. One of the problems with similarity and distance measurement between data dispersed with an N-dimensional space, where the magnitude of N is significant, is the computational complexity of calculating such measures. Much work has therefore been undertaken to attempt to reduce this complexity.

In the absence of a session chair (Shusaku Tasumoto, whose flight had been delayed) we agreed that Ömer Eğecioğlu, the first speaker, might as well introduce himself; which he did with practiced ease. In his presentation Ömer (from the University of California) proposed a similarity algorithm founded on data described in terms of both a magnitude and a shape part. Ömer noted that data described in this manner is of course an approximation, however it does serve to reduce the number of dimensions to a manageable level while maintaining overall "distance information". Ömer's experiments on time series data demonstrated that the approximation achieved good results while at the same time reducing the overall complexity of the input data. Vladimir Estivill-Castro, the second speaker, kindly acted as chair during the "any questions?" at the end of Ömer's talk (who needs session chairs!).

For the second presentation Vladimir Estivill-Castro introduced himself (University of Newcastle in Australia although, as I understand it, he is currently working in Canada). Vladimir commenced his presentation, on similar lines to speakers in the previous session, by observing that "clustering is in the eye of the beholder" – that is why there are so many different clustering approaches and why these

approaches are difficult to compare. Vladimir then went on to suggest that clustering can be improved by optimising the number of clusters. He illustrated this by considering bi-dimensional point data and described a technique founded on a quad-tree data structure of the form used in many geographic information systems. The advantage of this data structure is that distance information is implied by the structure of the tree itself, which can then be utilised to arrive at appropriate distance measures. (Shusaku, slightly flustered, arrived in time to chair the questions).

Having completed his first presentation (much to the confusion of the newly arrived session chair) Vladimir launched into his second presentation of the day. This concerned web mining, particularly measuring the similarity between user paths. Vladimir described an approach to clustering of such paths using a non crisp-classification approach that allows data to "have a degree of membership" with respect to a particular cluster (data can thus belong to several clusters). Some experimental output offered by Vladimir indicated encouraging results. This was the last presentation in the session and we headed off for coffee.

4 Sessions 5 and 6

The final sessions of the day comprised a "Logic-based approaches" session and a "Miscellaneous" session. I opted for the first. The first two papers in this session were concerned with the mining of deductive databases, and both made use of the concept of aggregates, i.e. functions that reduce a set of database records to some representational value (or values).

The first presentation was by Arno Knobbe of Kiminkii in the Netherlands and concentrated on addressing the difficulties of applying data mining techniques to relational databases comprising many tables. Arno proposed a solution whereby a single table is first constructed using aggregates which is then mined. The approach was illustrated using a number of data sets.

The next talk was by Fosca Giannotti of CNUCE-CNR, Italy. The talk concerned the embedding of data mining techniques within deductive query systems. Such systems are founded on query languages and consequently Fosca proposed the use of a DMQL (Data Mining Query Language). The problem then is how to specify an association rule mining algorithm within such a language given the complexity of such algorithms. Fosca proposed the use of iterative aggregates to support appropriate definition and illustrated the approach using the classic Apriori algorithm (Agrawal & Srikant, 1994).

The final talk was by Sergei Kuznetsov of the All-Russia Institute of Scientific and Technical Information (VINITI), and Sergei Obiedkov of the Russian State University (Moscow). The speakers considered the construction of concept lattices. Concept lattices are a useful method of hierarchically indicating the connection between attributes in a data set (some authors use the term *taxonomies* to describe such lattices). The two Sergeis presented a survey of concept lattice generation algorithms using a classification of such algorithms according to whether they operate in an incremental or batch manner. They then presented a series of time complexity graphs for a number of salient concept lattice algorithms. The conclusion of the work (disappointingly but to be expected) was that no one algorithm was optimum, but that the choice of algorithm should be influenced by the overall size and density of the input data.

5 Day Two

Day two opened with an invited talk by Antony Unwin of Augsburg University. There were two strands to the presentation. Antony commenced by considering the extraction of information from visual data sources and the problems encountered – essentially difficulties in understanding the graphics. Antony then went on to discuss some flawed data analysis studies and suggested that "statisfication" (application of statistical know how) would have benefitted these studies. The message was that a statistical understanding is required to "get the analysis right" and to identify the appropriateness of data sets. Antony also proposed that, more generally, statisification will also assist in the validation of data mining results.

After the invited talk we had coffee and then split into parallel streams and tutorial and workshop sessions. The tutorials for the day were on "Support vector machines" and "Co-training and learning from labeled and unlabeled data". Support Vector Machines (SVMs) (Cristianini & Shawe-Taylor, 2001) are a learning system based on statistical learning theory, and are currently considered to be the mathematical models (together with neural nets) that underlie learning. Recently there has been much interest within the machine learning community regarding SVMs, as witnessed by the special issue of *Machine Learning* on SVMs and kernel methods that came out in 2001. The workshops were on "Visual data mining", "Integrating aspects of data mining, decision support and meta-learning" and "Active learning, database sampling, experiment design: views on instance selection". Later in the day Michael Schroeder, of City University (London) told me that the visual data mining workshop was particularly good, however I elected to stick with the parallel sessions.

6 Sessions 7 and 8

The morning session included the first of two association rule mining sessions ("Fuzzy logic and association rules") and a session on text mining. As is often the case, I would have liked to go to both but opted for the "Fuzzy logic and association rules" session. The concept of fuzziness is an important issue in the context of association rule mining because, given a continuously valued (non-discrete) attribute we must partition the value range into a sequence of discrete "sub-ranges", each represented by a column in the revised data set (standard association rule mining algorithms work with only binary valued data) – this means we must derive cut-off points for our sub-ranges. However, instead of having distinct boundaries between sub-ranges it seems sensible to fuzzify the boundaries so that more than one sub-range may be applicable given a particular value. (Fuzziness also has a role to play when considering clustering partitions).

The first presentation was by Walter Kosters of Leiden University (Netherlands) who described a method of finding interesting rules using fuzzy techniques (as described above) and taxonomies (concept lattices) within which the notion of "lifting to parents" is used.

The second presentation was by Attila Gyenesai of the University of Turku in Finland. The subject of Attila's talk was the need to identify the upper and lower bounds of each fuzzy set. To this end a number of interestingness measures for fuzzy association rule mining were considered and evaluated: e.g. covariance and correlation measures, entropy measures, etc. The results appeared to be somewhat inconclusive in that the ranking of rules using different measures did not seem to be significantly different.

The third paper was by Eyke Hüllermeier who described a new approach to fuzzy association rule mining by modelling fuzzy associations as implication-based rules (as opposed to the more traditional approach based on set theory). Eyke then went on to consider various ways of interpreting fuzzy association rules, in turn suggesting quality measures and different approaches to generating such rules.

The final paper in this session on fuzzy logic and association rules was on fuzzy classification rules and was presented by Alex Freitas of PPGIA (Brazil). Alex described a co-evolutionary fuzzy classification system based on a genetic programming algorithm coupled with an evolutionary strategy algorithm. The results presented by Alex, comparing his algorithm with other evolutionary approaches, showed that the proposed system outperformed existing systems without encountering significant overheads.

I found all four papers in this session to be extremely worthwhile, and therefore considered this to be the most interesting session so far.

7 Sessions 9 and 10

After lunch it was time for our own paper in Session 9 on association rules (the parallel session, session 10, was on time series analysis). After this was excellently presented by Paul Leng, the second paper

was by Tobias Scheffer of the University of Magdeburg who described a Bayesian framework to determine the expected accuracy of association rule mining on future data. Tobias then went on to describe a "predicate Apriori" algorithm which implemented the approach. He completed his presentation with some observations on anomalies found in association rules such as redundancy and subsumption (anomalies not unlike those that rulebase/expert system V&V practioners attempt to discover and react to (Vermesan & Coenen, 1999)).

The final presentation in the session was by Heike Hofmann of AT&T research, USA. Heike commenced by considering that one of the problems found in association rule mining is that a very large number of rules is often generated (in some cases so large that KDD techniques should be applied to the list of rules). To address this problem Heike described some work (carried out in collaboration with, amongst others, Antony Unwin, who gave the keynote presentation in the morning) to visualise the result. She described a graphical technique, "the two-key plot" (the two keys being support and confidence) to display associations. She completed the talk by considering some aspects of rule pruning, and some results taken from an ornithological study carried out in Scotland. This completed the first of the afternoon sessions.

8 Sessions 11 and 12

After the tea break, I opted for Session 12, "Medical Applications" (Session 11 was on Web Mining and Collaborative Filtering). I went for the medical applications session, not because of any special interest in such applications, but because I wished to hear the presentation by Shusaku Tsumoto on time series analysis (having been unable to attend Session 10 which was devoted to the subject).

Shusaku Tsumoto, from the Shimane Medical University in Japan, gave a preliminary report on a mechanism for mining large temporal databases, specifically medical time series databases. The mechanism was founded on a rule induction approach. Shusaku concluded with some interesting results produced using the technique, indicating that a worthwhile end goal could be expected.

The second presentation, by Felice Roberts from Marquette University (USA), was concerned with machine-learning techniques to classify anomalies in ECG (electro cardiogram) recordings.

The final paper was a second presentation by Shusaku Tsumoto. This presentation was again concerned with rule induction and the mining of positive and negative association rules in clinical databases. The intuition here was that such rules model more closely the rules suggested by medical experts. Again the approach was supported by results from clinical data sets.

This ended day two of the conference. In the evening there was a reception in the "Kaufhaus" at which a representative of the mayor gave a welcome speech and we were entertained by members of the "academic orchestra of Freiburg".

9 Day Three

On Wednesday (day three) the two conferences (PKDD and ECML) came together, Wednesday marked the last day the of PKDD and start of ECML. The day was billed as the "joint conference" day and was commenced by a welcome by a representative of the University. Looking round the auditorium I thought that there were some 400 people in the audience.

In the spirit of the joint day many of the presentations crossed over the boundary between KDD and machine learning from one direction or another. The first invited talk of the day was by Heikki Mannila of the Helsinki University of Technology. Heikki considered that data mining research fell into two strands, modelling for prediction and finding patterns. He then went on to consider how these strategies are distinct and how they can be brought together and described an approach founded on the concept of maximum entropy.

10 Session 13

For the joint day there were no parallel sessions (and no tutorials or workshops), so after the coffee break we returned for a series of presentations with a "foot in both camps".

The first of these was by Thomas Hofmann of Brown University, USA, whose presentation was entitled "learning what people (don't) want". To decide what people want, Thomas described a collaborative filtering technique that performs an automatic decomposition of user preferences. The presentation was set in the context of recommendation systems (a disturbing trend where computer systems are used to tell people what they should watch on television; a decision they, apparently, cannot make for themselves because of the number of channels that are now available).

The second presentation was by Frank Höppner, from the University of Applied Science in Emden (Germany), who described a technique for identifying patterns in time series data founded on Allen's interval calculus (Allen, 1983). He illustrated his talk with an example which analysed the output from barographs. The third presentation was given by Ran El-Yaniv from the Israel Institute of Technology, who presented the "iterative double clustering" algorithm. Essentially this is a meta soft-clustering algorithm, useful when working in high dimensional data spaces.

The last presentation in the morning session was by Takashi Washio, from Osaka University. Takashi described a method for extracting simultaneous equations from data models. This was "ongoing" work first described at ECAI in 1977 where Takashi's SSF (Simultaneous Structure Finder) was first described. Takashi's presentation comprised a description of an extended version of SSF to work with passively observed data. This ended the morning session.

After lunch there was one more invited presentation which effectively marked the end of the joint day. The presentation was by Tom Dietterich of Oregon State University, and was on reinforcement learning. Reinforcement learning can best be described in terms of an agent searching through some environment attempting to reach some end goal, during which the agent explores and "learns" about the environment. The approach described by Tom Dietterich was based on SVM technology (see above).

11 Business meeting

PKDD ended with a business meeting to consider feedback from this year's conference and where we were going next. The feedback included (as always) criticism of the registration fee. Luc De Raedt broadly agreed with this observation but indicated that this was largely due to the increasing difficulty in gaining sponsorship. I felt that criticism over the registration fee was unjustified in that it compared very favorably with similar conferences running in 2001. I have also had experience, running conferences in the UK, of the difficulty in obtaining sponsorship.

With respect to PKDD 2002 there were two contenders for the "privilege" of running the conference(s), one group from Poland and one from Finland. A representative from each group made a short presentation outlining their "bid". The distinctions between the two bids were that (a) the Polish group were likely to be able to get more sponsorship than the Finnish; however (b) the Finnish delegation would run both PKDD and ECML back-to-back as was done this year, while the Polish group were only interested in PKDD. When put to a vote the fact that the Finnish group would run both PKDD and ECML back to back "won the day". Personally I thought that both bids were well put together; however, I was in tune with the idea of keeping PKDD and ECML together.

One other issue of note that was discussed at the business meeting was the setting up of the European Knowledge Discovery Network of Excellence (KDNet for short). The URL is `http://www.kdnet.org/` (although at time of writing, October 2001, this WWW site was still under construction).

References

Agrawal, R and Srikant, R, 1994, "Fast algorithms for mining association rules" in *Proceedings of 20th VLDB Conference* 487–499.

Allen, JF, 1983, "Maintaining knowledge about temporal intervals" *Communications of the ACM* **26**(11) 832–843.

Cristianini, N and Shawe-Taylor, J, 2001, *An Introduction to Support Vector Machines (and other Kernel-Based Learning Methods*) Cambridge University Press.

De Raedt, L and Flach, P (eds), 2001, *Machine Learning* (Proc. ECML 2001) Springer.

De Raedt, L and Siebes A (eds), 2001, *Principals of Data Mining and Knowledge Discovery* (Proc. PKDD 2001) Springer.

Vermesan. A and Coenen, F (eds.), 1999, *Validation and Verification of Knowledge Based Systems: Theory, Tools and Practice* Kluwer Academic publishing.