Advanced Review

Data mining tools

Ralf Mikut* and Markus Reischl

The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult. This paper attempts to support the decision-making process by discussing the historical development and presenting a range of existing state-of-the-art data mining and related tools. Furthermore, we propose criteria for the tool categorization based on different user groups, data structures, data mining tasks and methods, visualization and interaction styles, import and export options for data and models, platforms, and license policies. These criteria are then used to classify data mining tools into nine different types. The typical characteristics of these types are explained and a selection of the most important tools is categorized. This paper is organized as follows: the first section *Historical Development and State-of-the-Art* highlights the historical development of data mining software until present; the criteria to compare data mining software are explained in the second section *Criteria for Comparing Data Mining Software*. The last section *Categorization of Data Mining Software into Different Types* proposes a categorization of data mining software and introduces typical software tools for the different types. *WIREs Data Mining Knowl Discov* 2011 1 431–443 DOI: 10.1002/widm.24

HISTORICAL DEVELOPMENT AND STATE-OF-THE-ART

Data mining has a long history, with strong roots in statistics, artificial intelligence, machine learning, and database research.[1,2] The word 'data mining' can be found relatively early, as in the article of Lovell,[3] published in the 1980s. Advancements in this field were accompanied by development of related software tools, starting with mainframe programs for statistical analysis in the early 1950s, and leading to a large variety of stand alone, client/server, and web-based software as today's service solution.

Following the original definition given in Ref 1, data mining is a step in the knowledge discovery from databases (KDD) process that consists of applying data analysis and discovery algorithms to produce a particular enumeration of patterns (or models) across the data. In that same article, KDD is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Sometimes, the wider KDD definition is used synonymously for data mining. This wider interpretation is especially popular in the context of software tools because most such tools support the complete KDD process and not just a single step.

Today, a large number of standard data mining methods are available, (see Refs 4 and 5 for detailed descriptions). From a historical perspective, these methods have different roots. One early group of methods was adopted from classical statistics: the focus was changed from the proof of known hypotheses to the generation of new hypotheses. Examples include methods from Bayesian decision theory, regression theory, and principal component analysis. Another group of methods stemmed from artificial intelligence - like decision trees, rule-based systems, and others. The term 'machine learning' includes methods such as support vector machines and artificial neural networks. There are several different and sometimes overlapping categorizations; for example, fuzzy logic, artificial neural networks, and evolutionary algorithms, which are summarized as computational intelligence.[6]

The typical life cycle of new data mining methods begins with theoretical papers based on in-house software prototypes, followed by public or on-demand software distribution of successful algorithms as research prototypes. Then, either special commercial or open source packages containing a family of similar algorithms are developed or the algorithms are integrated into existing open source or commercial packages. Many companies have tried to promote their own stand alone packages, but only

*Correspondence to: ralf.mikut.kit.edu

Karlsruhe Institute of Technology, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, GERMANY

DOI: 10.1002/widm.24

few have reached notable market shares. The life cycle of some data mining tools is remarkably short. Typical reasons include internal marketing decisions and acquisitions of specialized companies by larger ones, leading to a renaming and integration of product lines.

The largest commercial success stories resulted from the step-wise integration of data mining methods into established commercial statistical tools. Companies such as SPSS, founded in 1975 with precursors from 1968, or SAS, founded in 1976, have been offering statistical tools for mainframe computers since the 1970s. These tools were later adapted to personal computers and client/server solutions for larger customers. With the increasing popularity of data mining, algorithms such as artificial neural networks or decision trees were integrated in the main products and specialized data mining companies such as Integrated Solutions Ltd. (acquired in 1998 by SPSS) were acquired to obtain access to data mining tools such as Clementine. During these periods, renaming of tools and company mergers played an important role in history; for example, the tool Clementine (SPSS) was renamed as PASW Modeler, and is now available as IBM SPSS Modeler after the acquisition of SPSS by IBM in 2009. In general, tools of this statistical branch are now very popular for the user groups in business application and applied research.

Concurrently, many companies offering business intelligence products have integrated data mining solutions into their database products; one example is Oracle Data Mining (established in 2002). Many of these products are also a product of the acquisition and integration of specialized data mining companies.

In 2008, the worldwide market for business intelligence (i.e., software and maintenance fees) was 7.8 billion USD, including 1.5 billion USD in so-called 'advanced analytics', containing data mining and statistics.[7] This sector has grown 12.1% between 2007 and 2008, with large players including companies such as SAS (33.2%, tool: SAS Enterprise Miner), SPSS (14.3%, since 2009, an IBM company; tool: IBM SPSS Modeler), Microsoft (1.7%, tool: SQL Server Analysis Services), Teradata (1.5%, tool: Teradata Database, former name TeraMiner), and TIBCO (1.4%, tool: TIBCO Spotfire).

Open-source libraries have also become very popular since the 1990s. The most prominent example is Waikato Environment for Knowledge Analysis (WEKA), see Ref 8. WEKA started in 1994 as a C++ library, with its first public release in 1996. In 1999, it was completely rebuilt as a JAVA package; since that time, it has been regularly updated. In addition, WEKA components have been integrated in many other open-source tools such as Pentaho, RapidMiner, and KNIME.

A large group of research prototypes are based on script-oriented mathematical programs such as MATLAB (commercial) and R (open source). Such mathematical programs were not originally focused on data mining, but contain many useful mathematical and visualization functions that support the implementation of data mining algorithms. Recently, graphical user interfaces such as those utilized for R (e.g., Rattle) and Matlab (e.g., Gait-CAD, Established in 2006) can be used as integration packages (INT) for many single, open-source algorithms.

As the number of available tools continues to grow, the choice of one special tool becomes increasingly difficult for each potential user. This decision-making process can be supported by criteria for the categorization of data mining tools. Different categorizations of tools were proposed in Refs 9–12. The last two comprehensive criteria-based surveys date back to 1999, covering 43 software packages in Ref 9, and 2003, with 33 tools in Ref 12 (a regularly updated Excel table is available on request from the same author with 63 tools in 2009). In addition, smaller reviews have been published, containing 12 open-source tools,[13] eight noncommercial tools,[14] nine commercial tools,[10] and five commercial tools using benchmark datasets.[15]

In the past 10–15 years, data mining has become a technology in its own right, is well established also in business intelligence (BI), and continues to exhibit steadily increasing importance in technology and life sciences sectors. For example, data mining was a key factor supporting methodological breakthroughs in genetics.[16] It is a promising technology for future fields such as text mining and semantic search engines,[17] learning in autonomous systems—as with humanoid robots[18] and cars, chemoinformatics[19] and others.

Various standardization initiatives have been introduced for data mining processes, data and model interfaces—as with Cross Industry Standard Process for Data Mining for industrial data mining,[20] and approaches focused on clinical and biological applications.[21] A survey of such initiatives is provided in Ref 22, and a large variety of standard data mining methods are described in comprehensive standard text books;[4, 5] however, new methods, especially for data streams,[23] extremely large datasets, graph mining,[24, 25] text mining,[17] and others have been proposed in the last few years. In the near future, methods for high-dimensional problems such as image retrieval[26] and video mining[27] will also be optimized and embedded into powerful tools.

TABLE 1 | Maximum Dimensions of Datasets for Different Types of Problems

Data	Dim.	Structure for Each of the *N* Examples
Feature table	2	*s* features (e.g., age and income)
Texts	2	frequency of words or *n*-grams (vector-space approach)
Time series	3	*s* time series with *K* time samples
Sequences	3	*s* sequences of length *L* (e.g., mass spectrograms and genes)
Images	4	*s* images with pixels
Graphs	4	*s* graphs with adjacency matrixes
3D images	5	*s* images with pixels and slices
Videos	5	*s* videos containing images with pixels and *K* time samples
3D videos	6	like videos, but with additional slices

Dim., maximum dimensionality; *s*, number of features; *N*, number of examples; *K*, number of samples in a time series. Lower dimensions of the dataset can occur for problems with only one feature $s = 1$ resp. one example ($N = 1$).

CRITERIA FOR COMPARING DATA MINING SOFTWARE

Survey

In the following, different criteria for comparison of data mining software are introduced. These criteria are based on user groups, data structures, data mining tasks and methods, import and export options, and license models. A detailed overview about the different tools is given later in this paper and as an Excel table in the additional material; however, some specific information about tools is discussed if a specific tool is unique to some aspects of the proposed criteria. The complete list of tools is provided toward the end of this paper.

User Groups

There are many different data mining tools available, which fit the needs of quite different user groups:

- Business applications: This group uses data mining as a tool for solving commercially relevant business applications such as customer relationship management, fraud detection, and so on. This field is mainly covered by a variety of commercial tools providing support for databases with large datasets, and deep integration in the company's workflow.
- Applied research: A user group that applies data mining to research problems, for example, technology and life sciences. Here, users are mainly interested in tools with well-proven methods, a graphical user interface (GUI), and interfaces to domain-related data formats or databases.
- Algorithm development: Develops new data mining algorithms, and requires tools to both integrate its own methods and compare these with existing methods. The necessary tools should contain many concurrent algorithms.
- Education: For education at universities, data mining tools should be very intuitive, with a comfortable interactive user interface, and inexpensive. In addition, they should allow the integration of in-house methods during programming seminars.

Data Structures

An important criterion is the dimensionality of the underlying raw data in the processed dataset (Table 1). The first data mining applications were focused on handling datasets represented as two-dimensional feature tables. In this classical format, a dataset consists of a set of *N* examples (e.g., clients of an insurance company) with *s* features containing real values or usually integer-coded classes or symbols (e.g., income, age, number of contracts, and alike). This format is supported by nearly all existing tools. In some cases, the dataset can be sparse, with only a few nonzero features such as a list of *s* shopping items for *N* different customers. The computational and memory effort can be reduced if a tool exploits this sparse structure.

Some structured datasets are characterized by the same dimensionality. As an example, sample documents in most text mining problems are represented by the frequency of words or so-called *n*-grams (a group of *n* subsequent characters in a document).[28]

The most prominent format having a higher dimensionality contains time series as elements, leading to dataset dimensions between one (i.e., only one example of a time series with *K* samples) and three (i.e., *N* different examples of *s*-dimensional vector time series with *K* samples). Typical tasks are forecasting of

future values, finding typical patterns in a time series or finding similar time series by clustering. The analysis of time series plays an import role in many different applications, including prediction of stock markets, forecasting of energy consumption and other markets, and quality supervision in production, and is also supported by most data mining tools.

With a similar dimensionality, different kinds of structured data exist such as gene sequences (spatial structure), spectrograms or mass spectrograms (structured by frequencies or masses), and others. Only a few tools support these types of structured data explicitly, but some tools for time series analysis can be rearranged to cope with these problems.

A more recent trend is the application of data mining methods for images and videos.[26, 27] The main challenge is the handling of extremely large raw datasets, up to gigabytes and terabytes, caused by the high dimensionality of the examples. Typical applications are microscopic images in biology and medicine, camera-based sensors in quality control and robotics, biometrics, and security. Such datasets must be split into metadata—with links to image and video files handled in a main dataset and files—which contain the main part of the data. Until now, these problems were normally solved using a combination of tools: the initial tool (e.g., ImageJ and ITK) would process the images or videos, resulting in segmented images and extracted features describing the segments; a second tool would solve data mining problems handling the extracted features as a classical table or time series.

Another format leading to image-like dimensions includes graphs that can be represented as adjacency matrices, describing the connection between different nodes of a graph. Graph mining has powerful applications,[24, 25] such as characterizing social networks and chemical structures; however, only a few such tools exist, including Pegasus and Proximity.

Tasks and Methods

The most important tasks in data mining are

- supervised learning, with a known output variable in the dataset, including
  - (a) classification: class prediction, with the variable typically coded as an integer output;
  - (b) fuzzy classification: with gradual memberships with values in-between 0 and 1 applied to the different classes;
  - (c) regression: prediction of a real-valued output variable, including special cases of predicting future values in a time series out of recent or past values;
- unsupervised learning, without a known output variable in the dataset, including
  - (a) clustering: finds and describes groups of similar examples in the data using crisp of fuzzy clustering algorithms;
  - (b) association learning: finds typical groups of items that occur frequently together in examples;
- semisupervised learning, whereby the output variable is known only for some examples.

Each of these tasks consists of a chain of low-level tasks. Furthermore, some low-level tasks can act as stand-alone tasks; for example, by identifying in a large dataset elements that possess a high similarity to a given example. Examples of such low-level tasks are:

- data cleaning (e.g., outlier detection);
- data filtering (e.g., smoothing of time series);
- feature extraction from time series, images, videos, and graphs (e.g., consisting of segmentation and segment description for images, characteristic values such as community structures in graphs);
- feature transformation (e.g., mathematical operations, including logarithms, dimension reduction by linear or nonlinear combinations by a principal component analysis, factor analysis or independent component analysis);
- feature evaluation and selection (e.g., by filter or wrapper methods);
- computation of similarities and detection of the most similar elements in terms of examples or features (e.g., by k-nearest-neighbor-methods and correlation analysis);
- model validation (cross validation, bootstrapping, statistical relevance tests and complexity measures);
- model fusion (mixture of experts); and
- model optimization (e.g., by evolutionary algorithms).

For almost all of these tasks, a large variety of classical statistical methods—including classifiers

using estimated probability density functions, factor analysis and others, and newer machine learning methods—such as artificial neural networks, fuzzy models, rough sets, support vector machines, decision trees, and random forests, are available. In addition, optimization models such as evolutionary algorithms can assist with the identification of model structures and parameters. The related methods are described in survey articles[29] or textbooks[4, 5] and are not summarized in this paper.

Not all of the data mining methods are available in all software tools. The following list contains a subjective evaluation of the frequency with which specific methods are incorporated in the different tools:

- Frequent: classifiers using estimated probability density functions, correlation analysis, statistical feature selection, and relevance tests;
- In many tools: decision trees, clustering, regression, data cleaning, data filtering, feature extraction, principal component analysis, factor analysis, advanced feature evaluation and selection, computation of similarities, artificial neural networks, model cross validation, and statistical relevance tests;
- In some tools: fuzzy classification, association learning and mining frequent item sets, independent component analysis, bootstrapping, complexity measures, model fusion, support vector machines, k-nearest-neighbor-methods, Bayesian networks, and learning of crisp rules;
- Infrequent: random forests[30] (contained in Waffles, Random Forests, WEKA, and all of its derivatives), learning of fuzzy systems (contained in KnowledgeMiner, See5, and Gait-CAD), rough sets[31] (in ROSETTA, and Rseslibs), and model optimization by evolutionary algorithms[14] (in KEEL, ADaM, and D2K).

Interaction and Visualization

There are three main types of interaction between a user and a data mining tool:

- pure textual interface using a programming language—difficult to handle, but easily automated;
- graphical interface with a menu structure—easy to handle, but not so easily automated; and
- graphical user interface where the user selects 'function blocks' or algorithms from a palette of choices, defines parameters, places them in a work area, and connects them to create complete data mining streams or workflows; a good compromise, but difficult to handle for large workflows.

Mixtures of these forms arise if macros of menu items can be recorded for workflows or if additional blocks in a workflow can be implemented using a programming language. Automation (scripting) is extremely important for routine tasks, especially with large datasets, because the workload of the user is reduced. Almost all tools provide powerful visualization techniques for the presentation of data mining results; particularly tools for business application and applied research, which are able to generate complete reports containing the most important results in a readable form for users lacking explicit data mining skills. Interactive methods can support an explorative data analysis. An example is a method called brushing that enables the user to select specific data points in a figure or subsets of data (e.g., nodes of a decision tree) and highlight these data points in other plots.

Import and Export of Data and Models

The ease with which data and models can be imported and exported among different software tools plays a crucial role in the functionality of data mining tools. First, the data are normally generated and hosted from different sources such as databases or software associated with measurement devices. In business applications, interfaces to databases such as Oracle or any database supporting the Structured Query Language (SQL) standard are the most common means of importing data. Because almost all other nondata mining tools support export as text or excel files, formats such as CSV (comma separated values) are frequently used to import formats with data mining tools. In addition, almost all software have proprietary binary or textual files, and exchanges formats for data and models, e.g., Attribute-Relation File Format in WEKA (WEKA standard).

In order to import and export developed models as components in other processes and systems, the XML-based standard PMML[32] was developed by the Data Mining Group (http://www.dmg.org) and is supported by many companies such as IBM and SAS. Another standard initiative is the Object Linking and Embedding Database (OLEDB, sometimes written as OLEDB or OLE-DB) for data mining, an API (Application Programming Interface) designed

by Microsoft to access different types of data stored in a uniform manner (http://msdn.microsoft.com/en-us/library/ms146608.aspx). OLEDB is a set of interfaces implemented using the Component Object Model (COM). For data exchange among different tools, another initiative deals with Java Specification Requests for data mining: versions 1.0 (JSR 73, final release in 2004: http://www.jcp.org/en/jsr/detail?id=73) and 2.0 (JSR 247, public review as last activity in 2006: http://www.jcp.org/en/jsr/detail?id=247) define an extensible Java API for data mining systems. The consortium includes many related companies, such as Oracle, SAS, SPSS (now IBM), SAP, and others; recent overviews can be found in Refs 33 and 34. Another interesting feature is the export of an executable runtime version of developed models. Often, they do not require a more expensive development license and can be run free of charge, or at least with a cheaper runtime license.

Platforms

Data mining tools can be subdivided into standalone and client/server solutions. Client/server solutions dominate, especially in products designed for business users. They are available for different platforms, including Windows, MAC OS, Linux, or special mainframe supercomputers. There is a growing number of JAVA-based systems that are platform-independent for users in research and applied research.

Further expected trends are an increasing number of web interfaces providing data mining as SAAS (software as a service, with tools like Data Applied) and a stronger support of client/server-based data mining solutions on grids (tool ADaM, e.g., see, steps to a standardization in Ref 35); however, both trends have the potential risk of hurting privacy policies because the protection of data is difficult and many companies are very careful with sensitive data.

Licenses

There exists a wide variety of data mining tools with commercial and open-source licenses. This is particularly true in the business application user group, where commercial software is very attractive due to high software stability, good coupling with other commercial tools for data warehouses, included software maintenance, and the possibility of user training for sophisticated topics. For all other user groups, there is a strong trend toward open-source software, but different types of licenses exist for this (e.g., see, survey in Ref 36). The main advantages of open-source software are faster bug fixes and methodological improvements, potential for integration with other tools, the existence of developer and user communities, faster adoption of methods to other innovative applications, and the fair comparison of new data mining algorithms with alternative ones. These advantages attract mainly users of applied research, development, and education; however, open-source tools are beginning to migrate even into business user groups,[37] particularly when additional commercial services such as training or maintenance are offered (e.g., Pentaho).

The most popular type of open-source licenses is the GNU General Public License of the Free Software Foundation (GNU-GPL or GPL: http://www.fsf.org). It permits free redistribution, integration in other packages, and modification of the software as long as all subsequent users receive the same level of freedom (so-called 'copy left'). This restriction guarantees that all software containing GNU-GPL components must be licensed under GNU-GPL. Weaker forms are licenses that are free for academic use, but not for business users.

Mixed forms of licenses occur especially if open-source software is used to expand commercial tools such as Matlab.

The Excel table (see, Section Supplementary Information) lists 195 recent tools (119 commercial tools, 67 open source tools, and nine tools with mixed license models).

CATEGORIZATION OF DATA MINING SOFTWARE INTO DIFFERENT TYPES

Following the criteria from the previous section, different types of similar data mining tools can be found. The typical characteristics of these types are explained in this section. Matching of the different types and user groups and the number of recent tools are summarized in Table 2. In addition, for commercial data mining tools, related tools and their group membership are summarized in different tables for commercial (Tables 3 and 4), free, and open-source data mining tools (Table 5). In these tables, very popular tools are marked in bold. The popularity was measured by

- the 20 most frequently used tools for real projects from 'Data Mining/Analytic Tools Used Poll 2010' of KDnuggets with 912 voters (http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html); [top 10 tools were RapidMiner, R, Excel (here ignored), KNIME, Pentaho/WEKA, SAS, MATLAB,

TABLE 2 | Matching Between Different User Groups and Tool Types with Number of Recent Tools in the Excel Table (see, Section Supplementary Information, tools belonging to two categories are counted twice)

Types	Data Mining Suites	Business Intelligence Packages	Mathematical Packages	Integration Packages	Extensions	Data Mining Libraries	Specialities	Research Prototypes	Solutions
Number of Recent Tools	46	16	5	8	10	20	56	17	19
Business applications	+	+	−	0	0	−	0	−	0
Applied research	+	−	+	+	0	0	0	0	+
Algorithm development	−	−	+	+	−	+	0	+	−
Education	+	−	0	0	+	−	+	−	0

Evaluation, +: especially useful, 0: less useful, −: not useful.

IBM SPSS Statistics, IBM SPSS Modeler, and Microsoft SQL Server];

- all main products of vendors with more than 1% market share in the section 'Advanced Analytics Tools' from Ref 7; and
- the most popular image processing tools (ITK and ImageJ) from the author's own experience to cover this field.

In this paper, the following nine types are proposed:

- **Data mining suites (DMS)** focus largely on data mining and include numerous methods. They support feature tables and time series, while additional tools for text mining are sometime available. The application focus is wide and not restricted to a special application field, such as business applications; however, coupling to business solutions, import and export of models, reporting, and a variety of different platforms are nonetheless supported. In addition, the producers provide services for adaptation of the tools to the workflows and data structures of the customer. DMS is mostly commercial and rather expensive, but some open-source tools such as RapidMiner exist. Typical examples include IBM SPSS Modeler, SAS Enterprise Miner, Alice d'Isoft, DataEngine, DataDetective, GhostMiner, Knowledge Studio, KXEN, NAG Data Mining Components, Partek Discovery Suite, STATISTICA, and TIBCO Spotfire.
- **Business intelligence packages (BIs)** have no special focus to data mining, but include basic data mining functionality, especially for statistical methods in business applications. BIs are often restricted to feature tables and time series, large feature tables are supported. They have a highly developed reporting functionality and good support for education, handling, and adaptation to the workflows of the customer. They are characterized by a strong focus on database coupling, and are implemented via a client/server architecture. Most BI softwares are commercial (IBM Cognos 8 BI, Oracle Data Mining, SAP Netweaver Business Warehouse, Teradata Database, DB2 Data Warehouse from IBM, and PolyVista), but a few open-source solutions exist (Pentaho).
- **Mathematical packages (MATs)** have no special focus on data mining, but provide a

TABLE 3 | List of Commercial Tools (Part 1)

Tool	Type	Link
ADAPA (Zementis)	DMS	www.zementis.com
Alice (d'Isoft)	DMS	www.alice-soft.com
Bayesia Lab	SPEC	www.bayesia.com
C5.0	SPEC	www.rulequest.com
CART	SPEC	www.salford-systems.com
Data Applied	DMS	data-applied.com
DataDetective	DMS	www.sentient.nl/?dden
DataEngine	DMS	www.dataengine.de
Datascope	DMS	www.cygron.hu
DB2 Data Warehouse	BI	www.ibm.com/software/data/infosphere/warehouse
DeltaMaster	BI	www.bissantz.com/deltamaster
Forecaster XL	EXT	www.alyuda.com
GhostMiner	DMS	www.fqs.pl/business_intelligence/products/ghostminer
IBM Cognos 8 BI	BI	www.ibm.com/software/data/cognos/data-mining-tools.html
IBM SPSS Modeler	DMS	www.spss.com/software/modeling/modeler
IBM SPSS Statistics	MAT	www.spss.com/software/statistics
iModel	DMS	www.biocompsystems.com/products/imodel
InfoSphere Warehouse	BI	www.ibm.com/software/data/infosphere/warehouse
JMP	DMS	www.jmpdiscovery.com
KnowledgeMiner	SPEC	www.knowledgeminer.net
KnowledgeStudio	DMS	www.angoss.com
KXEN	DMS	www.kxen.com
Magnum Opus	SPEC	www.giwebb.com
MATLAB	MAT	www.mathworks.com
MATLAB Neural Network Toolbox	EXT	www.mathworks.com
Model Builder	DMS	www.fico.com
ModelMAX	SOL	www.asacorp.com/products/mmxover.jsp

Very popular tools are marked in bold letters.

large and extendable set of algorithms and visualization routines. They support feature tables, time series, and have at least import formats for images. The user interaction often requires programming skills in a scripting language. MATs are attractive to users in algorithm development and applied research because data mining algorithms can be rapidly implemented, mostly in the form of extensions (EXT) and research prototypes (RES). MAT packages exist as commercial (MATLAB and R-PLUS) or open-source tools (R, Kepler). In principle, table calculation software such as Excel may also be categorized here, but it is not included in this paper. Most tools are available for different platforms but have weaknesses in database coupling.

- **Integration packages** (**INTs**) are extendable bundles of many different open-source algorithms, either as stand-alone software (mostly based on Java; as KNIME, the GUI-version of WEKA, KEEL, and TANAGRA) or as a kind of larger extension package for tools from the MAT type (such as Gait-CAD, PRTools for MATLAB, and RWEKA for R). Import and export support standard formats, but database support is quite weak. Most tools are available for different platforms and include a GUI. Mixtures of license models occur if open-source integration packages are based on commercial tools from the MAT type. With these characteristics, the tools are attractive to algorithm developers and users in applied research due to expandability and rapid comparison with alternative tools, and due to easy integration of application-specific methods and import options.
- **EXT** are smaller add-ons for other tools such as Excel, Matlab, R, and so forth, with limited but quite useful functionality. Here, only a few data mining algorithms are implemented

TABLE 4 | List of Commercial Tools (Part 2)

Tool	Type	Link
Molegro Data Modeler	SOL	www.molegro.com
NAG Data Mining Components	LIB	www.nag.co.uk/numeric/DR/DRdescription.asp
NeuralWorks Predict	SPEC	www.neuralware.com/products.jsp
Neurofusion	LIB	www.alyuda.com
Neuroshell	SPEC	www.neuroshell.com
Oracle Data Mining (ODM)	DMS	www.oracle.com/technology/products/bi/odm/index.html
Partek Discovery Suite	DMS	www.partek.com/software
Partek Genomics Suite	SOL	www.partek.com/software
PolyAnalyst	DMS	www.megaputer.com/polyanalyst.php
PolyVista	BI	www.polyvista.com
Random Forests	SPEC	www.salford-systems.com
RapAnalyst	SPEC	www.raptorinternational.com/rapanalyst.html
R-PLUS	MAT	www.experience-rplus.com
SAP Netweaver Business Warehouse (BW)	BI	www.sap.com/platform/netweaver/components/businesswarehouse
SAS Enterprise Miner	DMS	www.sas.com/products/miner
See5	SPEC	www.rulequest.com
SPAD Data Mining	DMS	eng.spadsoft.com
SQL Server Analysis Services	DMS	www.microsoft.com/sql
STATISTICA	DMS	www.statsoft.com/products/data-mining-solutions/G259
SuperQuery	DMS	www.azmy.com
Teradata Database	BI	www.teradata.com
Think Enterprise Data Miner (EDM)	DMS	www.thinkanalytics.com
TIBCO Spotfire	DMS	spotfire.tibco.com
Unica PredictiveInsight	DMS	www.unica.com
WizRule and WizWhy	SPEC	www.wizsoft.com
XAffinity	SPEC	www.exclusiveore.com

Very popular tools are marked in bold letters.

such as artificial neural networks for Excel (Forecaster XL and XLMiner) or MATLAB (Matlab Neural Networks Toolbox). There are commercial or open-source versions, but licenses for the basic tools must also be available. The user interaction is the same as for the basic tool, for example, by using a programming language (MATLAB) or by embedding the extension in the menu (Excel).

- **Data mining libraries (LIBs)** implement data mining methods as a bundle of functions. These functions can be embedded in other software tools using an Application Programming Interface (API) for the interaction between the software tool and the data mining functions. A graphical user interface is missing, but some functions can support the integration of specific visualization tools. They are often written in JAVA or C++ and the solutions are platform independent. Open source examples are WEKA (Java-based), MLC++ (C++ based), JAVA Data Mining Package, and LibSVM (C++ and JAVA-based) for support vector machines. A commercial example is Neurofusion for C++, whereas XELOPES (Java, C++, and C♯) uses different license models. LIB tools are mainly attractive to users in algorithm development and applied research, for embedding data mining software into larger data mining software tools or specific solutions for narrow applications.
- **Specialties (SPECs)** are similar to DMS tools, but implement only one special family of methods such as artificial neural networks. They contain many elaborate visualization techniques for such methods. SPECs are rather simple to handle as compared with other tools, which eases the use of such tools in education. Examples are CART for decision trees, Bayesia Lab for Bayesian networks, C5.0, WizRule, Rule Discovery System for rule-based systems, MagnumOpus for association analysis, and JavaNNS, Neuroshell,

TABLE 5 | List of Free and Open-Source Tools

Tool	Type	Link
ADaM*	LIB	datamining.itsc.uah.edu/adam
CellProfilerAnalyst	SOL	www.cellprofiler.org/index.htm
D2K*	DMS	alg.ncsa.uiuc.edu
Gait-CAD	INT	sourceforge.net/projects/gait-cad
GATE	SOL	gate.ac.uk/download
GIFT	RES	www.gnu.org/software/gift
Gnome Data Mine Tools	DMS	www.togaware.com/datamining/gdatamine
Himalaya	RES	himalaya-tools.sourceforge.net
ImageJ	SOL	rsbweb.nih.gov/ij
ITK	SOL	www.itk.org
JAVA Data Mining Package	LIB	sourceforge.net/projects/jdmp
JavaNNS	SPEC	www.ra.cs.uni-tuebingen.de/software/JavaNNS/welcome_e.html
KEEL	INT	www.keel.es
Kepler	MAT	kepler-project.org
KNIME	INT	www.knime.org
LibSVM	LIB	www.csie.ntu.edu.tw/ cjlin/libsvm
MEGA	SOL	www.megasoftware.net/m_distance.html
MLC++	LIB	www.sgi.com/tech/mlc
Orange	LIB	www.ailab.si/orange
Pegasus	RES	www.cs.cmu.edu/ pegasus
Pentaho	BI	sourceforge.net/projects/pentaho
Proximity	SPEC	kdl.cs.umass.edu/proximity/index.html
PRTools	EXT	www.prtools.org
R	MAT	www.r-project.org
RapidMiner	DMS	www.rapidminer.com
Rattle	INT	rattle.togaware.com
ROOT	LIB	root.cern.ch/root
ROSETTA	SPEC	www.lcb.uu.se/tools/rosetta/index.php
Rseslibs	RES	logic.mimuw.edu.pl/ rses
Rule Discovery System*	SPEC	www.compumine.com
RWEKA	INT	cran.r-project.org/web/packages/RWeka/index.html
TANAGRA	INT	eric.univ-lyon2.fr/ ricco/tanagra/en/tanagra.html
Waffles	LIB	waffles.sourceforge.net
WEKA	DMS, LIB	sourceforge.net/projects/weka
XELOPES Library*	LIB	www.prudsys.de/en/technology/xelopes
XLMiner*	EXT	www.resample.com/xlminer

Very popular tools are marked in bold letters.
*, Commercial tools with free licenses for academic use.

NeuralWorks Predict, RapAnalyst for artificial neural networks.

- **RES** are usually the first—and not always stable—implementations of new and innovative algorithms. They contain only one or a few algorithms with restricted graphical support and without automation support. Import and export functionality is rather restricted and database coupling is missing or weak. RES tools are mostly opensource. They are mainly attractive to users in algorithm development and applied research, specifically in very innovative fields. Examples are GIFT for content-based image retrieval, Himalaya for mining maximal frequent item sets, sequential pattern mining and scalable linear regression trees, Rseslibs for rough sets, and Pegasus for graph mining. Early versions of today's popular tools such as WEKA and RapidMiner started in this category and shifted later to other categories as DMS.
- **Solutions** (**SOLs**) describe a group of tools that are customized to narrow application fields such as text mining (GATE), image

processing (ITK, ImageJ), drug discovery (Molegro Data Modeler), image analysis in microscopy (CellProfilerAnalyst), or mining gene expression profiles (Partek Genomics Suite, MEGA). The advantage of these solutions is the excellent support of domain-specific feature extraction techniques, evaluation measures, visualizations, and import formats. The level of data mining methods ranges from rather weak support (particularly in image processing) to highly developed algorithms. In some cases, more general tools from types DMS or INT also support specific domains (KNIME, Gait-CAD for peptide chemoinformatics). There are many commercial and open-source solutions.

A large variety of tools actually requires a fuzzy categorization with gradual memberships to different types. Examples are tools including a set of different algorithms (LIB) with an additional GUI acting as an INT, DMS, including special methods for narrow application fields and others. In these cases, a main type was assigned and the other fuzzy memberships are discussed in the Excel table in the additional material section.

The following kinds of tools were not included in the comparison:

- nonavailable software (e.g., owing to company mergers or stopped developments) is only listed in the Excel table in the additional material,
- software for the handling of data warehouses without explicit focus on data mining,
- software for the manual design and application of rule-based systems,
- software for table calculation with a focus to office users, and
- customized solutions for very narrow fields.

CONCLUSION

Many advanced tools for data mining are available either as open-source or commercial software. They cover a wide range of software products, from comfortable problem-independent data mining suites, to business-centered data warehouses with integrated data mining capabilities, to early research prototypes for newly developed methods. In this paper, nine different types of tools are presented: DMS, BIs, MATs, INT, EXT, SPECs, RES, LIBs, and SOLs. They vary in many different characteristics, such as intended user groups, possible data structures, implemented tasks and methods, interaction styles, import and export capabilities, platforms and license policies are variable. Recent tools are able to handle large datasets with single features, time series, and even unstructured data-like texts; however, there is a lack of powerful and generalized mining tools for multidimensional datasets such as images and videos.

SUPPLEMENTARY INFORMATION

An additional Excel table contains a list of 269 tools (195 recent and 74 historical tools, version from July 22, 2010). For each tool, the following information is available:

- toolbox name,
- company or group (with the term 'various' for open-source projects without an explicit developer),
- categorization into types with abbreviations for Research Prototypes (RES), Data Mining Libraries (LIB), Business Intelligence Packages (BI), Data Mining Software (DMS), Specialties (SPEC), Mathematical Packages (MAT), Extensions (EXT), Integration Packages (INT), Solutions (SOL),
- Giraud-Carrier: marking the covering by the Excel table in Ref 12 (Stand: February 3, 2010) with the values 1 (included in a detailed categorization), −1 (excluded), empty field: not mentioned,
- remarks,
- web link,
- activity: 1 (relevant tool, included in the comparison), 0 (less relevant), −1 (not available).
- license: OS, open source; CO, commercial; CO/OS, different versions available.

There are a number of regularly updated web resources with link lists, but lacking a criteria-based comparison of the tools. The most important web resources are:

- KDnuggets: *http://www.kdnuggets.com/software/suites.html*, including regular polls to identify the most frequently used tools,

- The Data Mine: *http://www.the-data-mine.com/bin/view/software*
- The Open Directory Project: *http://www.dmoz.org/Computers/Software/Databases/Data_Mining*
- Sourceforge (very popular platform for open-source solutions, search for 'data mining' to find data mining tools hosted at Sourceforge): *http://sourceforge.net/*
- Kernel Machines (especially to get a list of software to support vector machines): *http://www.kernel-machines.org/software*
- Tools for Bayesian Networks: *www.cs.helsinki.fi/research/cosco/Bnets*.

ACKNOWLEDGMENTS

The authors thank C. Giraud-Carrier for a copy of an Excel table containing a large set of data mining tools, the anonymous reviewers for many comments and suggestions, and R. A. Klady for the critical proofreading of the manuscript.

REFERENCES

1. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Mag* 1996, 17:37–54.
2. Smyth P. Data mining: Data analysis on a grand scale? *Stat Methods Med Res* 2000, 9:309–327.
3. Lovell MC. Data mining. *Rev Econ Stat* 1983, 65:1–11.
4. Han J, Kamber M. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann; 2006.
5. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer; 2008.
6. Engelbrecht AP. *Computational Intelligence - An Introduction*. Chichester: John Wiley; 2007.
7. Vesset D, McDonough B. Worldwide business intelligence tools 2008 vendor shares, IDC Competitive Analysis Report (2009).
8. Frank E, Hall M, Holmes G, Kirkby R, Pfahringer B, Witten I. Weka: A machine learning workbench for data mining. *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. New York: Springer; 2005, 1305–1314.
9. Goebel M. A survey of data mining and knowledge discovery software tools, ACM SIGKDD Explorations. *Newsletter* 1999, 1:20–33.
10. Wang J, Hu X, Hollister K, Zhu D. A comparison and scenario analysis of leading data mining software. *Int J Knowl Manage* 2008, 4:17–34.
11. Wang J, Chen Q, Yao J. Data mining software. In: Tomei L, ed., *Encyclopedia of Information Technology Curriculum Integration*. Hershey, PA: Information Science Publishing; 2008, 173–178.
12. Giraud-Carrier C, Povel O. Characterising data mining software. *Intell Data Anal* 2003, 7:181–192.
13. Chen X, Ye Y, Williams G, Xu X. A survey of open source data mining systems, Lecture Notes in Computer Science 2007, 4819:3–14.
14. Alcalá-Fdez J, Sánchez L, García S, del Jesus M, Ventura S, Garrell J, Otero J, Romero C, Bacardit J, Rivas V, et al. KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Comput* 2009, 13:307–318.
15. Haughton D, Deichmann J, Eshghi A, Sayek S, Teebagy N, Topi H. A review of software packages for data mining. *Am Stat* 2003, 57:290–310.
16. Barrett T, Troup D, Wilhite S, Ledoux P, Rudnev D, Evangelista C, Kim I, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: Mining tens of millions of expression profiles–database and tools update. *Nucleic acids Res* 2007, D760.
17. Weiss S. Text mining: predictive methods for analyzing unstructured information. New York: Springer-Verlag; 2005.
18. Dillmann R. Teaching and learning of robot tasks via observation of human performance. *Rob Auton Syst* 2004, 47:109–116.
19. Leach A, Gillet V. *An Introduction to Chemoinformatics*. Springer; 2007.
20. Shearer C. The CRISP-DM model: The new blueprint for data mining. *J Data Warehousing* 2000, 5:13–22.
21. Mikut R, Reischl M, Burmeister O, Loose T. Data mining in medical time series. *Biomed Tech* 2006, 51:288–293.

22. Grossman R, Hornick M, Meyer G. Data mining standards initiatives. *Commun ACM* 2002, 45:61.
23. Muthukrishnan S. *Data Streams: Algorithms and Applications*. Hanover, MA: Now Publishers Inc.; 2005.
24. Chakrabarti D, Faloutsos C. Graph mining: laws, generators, and algorithms. *ACM Comput Surv (CSUR)* 2006, 38:1–69.
25. Borgelt C. Graph mining: An overview. Proc., 19. *Workshop Computational Intelligence*. Karlsruhe, Germany: KIT Scientific Publishing; 2009, 189–203.
26. Datta R, Joshi D, Li J, Wang J. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput Surv (CSUR)* 2008, 40:1–60.
27. Zhu X, Wu X, Elmagarmid A, Feng Z, Wu L. Video data mining: Semantic indexing and event detection from the association perspective. *IEEE Trans Knowl Data Eng* 2005, 17:665–677.
28. Damashek M. Gauging similarity with n-Grams: Language-independent categorization of text. *Science* 1995, 267:843–848.
29. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: A review. *IEEE Trans Pattern Anal Mach Intell* 2000, 22:4–36.
30. Breiman L. Random forests. *Mach Learn* 2001, 45:5–32.
31. Pawlak Z. Rough sets and intelligent data analysis. *Inf Sci* 2002, 147:1–12.
32. Pechter R. What's PMML and what's new in PMML 4.0?, ACM SIGKDD Explorations. *Newsletter* 2009, 11:19–25.
33. Hornick M, Marcadé E, Venkayala S. Java Data Mining: Strategy, Standard, and Practice: A Practical Guide for Architecture, Design, and Implementation. San Francisco: Morgan Kaufmann Publishers Inc.; 2006.
34. Anand S, Grobelnik M, Herrmann F, Hornick M, Lingenfelder C, Rooney N, Wettschereck D. Knowledge discovery standards. *Artificial Intelligence Review* 2007, 27:21–56.
35. Cannataro M, Congiusta A, Pugliese A, Talia D, Trunfio P. Distributed data mining on grids: Services, tools, and applications. *IEEE Trans Syst Man Cybern B Cybern* 2004, 34:2451–2465.
36. Sonnenburg S, Braun M, Ong C, Bengio S, Bottou L, Holmes G, LeCun Y, Müller K, Pereira F, Rasmussen C, et al. The need for open source software in machine learning. *J Mach Learn Res* 2007, 8:2443–2466.
37. Bitterer A. Open-source business intelligence tool production deployments will grow five-fold through 2010, Gartner RAS Research Note G00171189 (2009).

