Knowledge Discovery Practices and Emerging Applications of Data Mining:

Trends and New Domains

A. V. Senthil Kumar
Hindusthan College of Arts and Science, Bharathiar University, India

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE

Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Book Publications: Julia Mosemann
Acquisitions Editor: Lindsay Johnston
Development Editor: Joel Gamon
Publishing Assistant: Milan Vracarich, Jr.
Typesetter: Natalie Pronio
Production Editor: Jamie Snavely
Cover Design: Lisa Tosheff

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Knowledge discovery practices and emerging applications of data mining : trends and new domains / Av Senthil Kumar, editor.
p. cm.
Includes bibliographical references and index. Summary: "This book introduces the reader to recent research activities in the field of data mining, covering association mining, classification, mobile marketing, opinion mining, microarray data mining, internet mining and applications of data mining on biological data, telecommunication and distributed databases"--Provided by publisher. ISBN 978-1-60960-067-9 (hardcover) -- ISBN 978-1-60960-069-3 (ebook) 1. Data mining. 2. Knowledge acquisition (Expert systems) I. Senthil kumar, A., 1966-
QA76.9.D343K 5645 2010
006.3'12--dc22
2010027734

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 11
A Comparative Study of Associative Classifiers in Mesenchymal Stem Cell Differentiation Analysis

Weiqi Wang
University of Oxford, UK

Yanbo J. Wang
China Minsheng Banking Corporation Ltd., China

Qin Xin
Simula Research Laboratory, Norway

René Bañares-Alcántara
University of Oxford, UK

Frans Coenen
University of Liverpool, UK

Zhanfeng Cui
University of Oxford, UK

ABSTRACT

Discovering how Mesenchymal Stem Cells (MSCs) can be differentiated is an important topic in stem cell therapy and tissue engineering. In a general context, such differentiation analysis can be modeled as a classification problem in data mining. Specifically, this is concerned with the single-label multi-class classification task. Previous studies on this topic suggests the Associative Classification (AC) rather than other alternative (Classification) techniques, and presented classification results based on the CMAR (Classification based on Multiple Association Rules) associative classifier. Other AC algorithms include: CBA (Classification Based on Associations), PRM (Predictive Rule Mining), CPAR (Classification based

DOI: 10.4018/978-1-60960-067-9.ch011

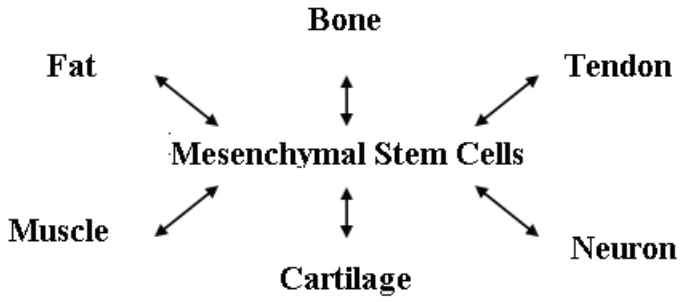
on Predictive Association Rules) and TFPC (Total From Partial Classification). The main aim of this chapter is to compare the performance of different associative classifiers, in terms of classification accuracy, efficiency, number of rules to be generated, quality of such rules, and the maximum number of attributes in rule-antecedents, with respect to MSC differentiation analysis.

INTRODUCTION

Mesenchymal Stem Cells (MSCs) have been claimed to be an integral part of tissue engineering due to their pluripotent differentiation potential both *in vivo* and *in vitro* (Beeres, Atsma, van der Laarse, Pijnappels, van Tuyn, & Fibbe, 2005; Derubeis & Cancedda, 2004; Zhang, Li, Jiang, Wu, & Liu, 2004), and have become one of the most significant research topics in the past few decades. MSCs are able to differentiate along the osteogenic, chondrogenic, adipogenic, myogenic, tendonogenic, and neurogenic lineages under appropriate stimuli (Pittenger, Mackay, Beck, Jaiswal, Douglas, & Mosca, 1999; Roelen & Dijke, 2003; Tuan, Boland, & Tuli, 2003), generating bone, cartilage, fat, muscle, tendon, and neuron cells respectively (Figure 1). Other discoveries on plasticity and immunologic properties of MSCs have further increased the interest in their clinical applications (Krampera, Glennie, Dyson, Scott, Laylor, & Simpson, 2003; Muller, Kordowich, Holzwarth, Spano, Isenee, & Staiber, 2006). The significance of MSCs in clinical therapy has triggered an urgent need for a better understanding and, if possible, computational prediction of MSCs differentiation (Griffith & Swartz, 2006).

In order to obtain a better understanding of MSCs, a significant number of studies have been conducted (Battula, Bareiss, Treml, Conrad, Albert, & Hojak, 2007; Hanada, Dennis, & Caplan, 1997; Lennon, Haynesworth, Young, Dennis, & Caplan, 1995; Magaki, Kurisu, & Okazaki, 2005; Meuleman, Tondreau, Delforge, Dejeneffe, Massy, & Libertalis, 2006; Muller et al., 2006), providing an enormous amount of experimental data for computational prediction. However, those studies and experiments were not interrelated with each other, i.e. different experiments focused on different combinations of factors affecting MSC differentiation, including species of cell donors, *in vitro* vs. *in vivo* environments where the experiments were executed, cell culture media, growth factors and supplements to the culture media, culture dimension (monolayer vs. 3D culture), cell attaching substrate (for monolayer culture) vs. scaffold (for 3D culture), and cell behaviors, especially the differentiation fates of

Figure 1. Differentiation fates of MSCs



MSCs in terms of the different lineages to which the cells committed (Hanada et al., 1997; Haynesworth, Baber, & Caplan, 1996; Kuznetsov, Friedenstein, & Robey, 1997; Lennon et al., 1995; Muller et al., 2006). The scattered experimental data hence resulted in a large amount of noise in the database and a discrete data structure, which cannot take advantage of traditional mathematical modeling methods. As a consequence, it is extremely difficult to construct intracellular pathway models for MSC metabolism, especially for their differentiation process (Bianco, Riminucci, Gronthos, & Robey, 2001).

On the other hand, useful information and meaningful prediction for MSC differentiation can be derived based on knowledge discovery via data mining techniques. The nature of data mining is to discover useful, but hidden, information (knowledge) in data. Previous studies under this heading (Wang, Wang, Banares-Alcantara, Coenen, & Cu, in press; Wang, Wang, Banares-Alcantara, Cui, & Coenen, 2009) model the analysis of MSC differentiation as a classification problem (in data mining) — the task of assigning predefined categories (differentiation fates) to "unseen" (MSC) instances. Broadly speaking, classification can be separated into two divisions: *single-label* that assigns exactly one predefined category to each "unseen" instance; and *multi-label* that assigns one or more predefined category to each "unseen" instance. With regard to *single-label* classification, three distinct approaches can be identified: *one-class* which learns from positive data samples only, and either assigns the predefined category to a "unseen" instance or ignores the assignation of the instance; *two-class* (or *binary*) which learns from both positive and negative data samples, and assigns either a predefined category or the complement of this category to each "unseen" instance; and *multi-class* which simultaneously deals with all given categories comprising all data samples, and assigns the most appropriate category to each "unseen" instance. The study presented in this chapter is concerned with the *single-label multi-class* classification task.

Mechanisms on which classification algorithms have been based include: decision trees, naive Bayes, *k*-NN (*k*-Nearest Neighbor), SVM (Support Vector Machine), genetic algorithm, neural networks, inductive learners (such as *FOIL* (First Order Inductive Learner) and RIPPER (Repeated Incremental Pruning to Produce Error Reduction)), association rules, etc. Among these mechanisms, classification based on association rules, i.e. Associative Classification (AC) or Classification Association Rule Mining (CARM), was suggested to address the MSC differentiation analysis problem (Wang et al., in press; Wang et al., 2009). It seems that AC (or CARM) offers a number of advantages over other classification approaches (Coenen, Leng, & Zhang, 2005; Shidara, Nakamura, & Kudo, 2007; Thabtah, Cowling, & Peng, 2005).

Coenen and Leng (2007) indicate:

- "*Training of the classifier is generally much faster using CARM (AC) techniques than other classification generation techniques such as decision tree (induction) and SVM (support vector machine) approaches*" (particularly when handling with the *multi-class* problem).
- "*Training sets with high dimensionality can be handled very effectively*".
- "*The resulting classifier is expressed as a set of rules which are easily understandable and simple to apply to unseen data (an advantage also shared by some other techniques, e.g. decision tree classifiers)*".
- In addition Liu et al. (1998) suggest that "*Experimental results show that the classifier built this way (AC) is, in general, more accurate than that produced by the state-of-the-art classification system*".

Since the first introduction of AC (Ali, Manganaris, & Srikant,, 1997), a number of major (AC) algorithms have emerged, these include: CBA (Classification Based on Associations), CMAR (Classification based on Multiple Association Rules), PRM (Predictive Rule Mining), CPAR (Classification based on Predictive Association Rules), and TFPC (Total From Partial Classification). Broadly speaking, these AC algorithms can be categorized into two groups, described as follows, according to the way that the Classification Association Rules (CARs) are generated.

- **Two Stage Algorithms,** were a set of CARs are produced first (as "stage 1"), which are then pruned and placed into a classifier (as "stage 2"). Typical algorithms of this approach include CBA (Liu et al., 1998) and CMAR (Li et al., 2001).
- **Integrated Algorithms,** were the classifier is produced in a single processing step. Algorithms of this kind include PRM and CPAR (Yin & Han, 2003), and TFPC (Coenen & Leng, 2004, 2007; Coenen et al., 2005).

Previous studies in data mining (or knowledge discovery) based MSC differentiation analysis report a satisfactory performance using the CMAR associative classifier, with regard to an online MSC database (Wang, Wang, Banares-Alcantara, Coenen, & Cui, in press; Wang, Wang, Banares-Alcantara, Cui, & Coenen, 2009). In this chapter, the analysis of MSC differentiation by addressing a series of AC approaches is developed, and aim to find the most appropriate associative classifier for this MSC differentiation study, by comparing the performance of different (AC) approaches in several aspects, i.e. classification accuracy, efficiency, number of rules to be generated, quality of such rules, and maximum number of attributes in rule-antecedents.

Chapter Organization

The rest of this chapter is organized as follows. The following section describes some related data mining aspects, as the background knowledge of this chapter, in classification, Association Rule Mining (Cody, Boctor, Filley, Hazen, Scott, & Sharma, 2000), and Associative Classification (AC). In the third section, five existing AC approaches (i.e. CBA, CMAR, PRM, CPAR and TFPC) are described in detail. The construction of a domain-specific (MSC) database, as the data preparation of the study, is introduced in the fourth section. Experiments are presented in the fifth section that compares the performance of existing AC approaches in MSC differentiation study. The sixth section gives a discussion of the study, and further points out some future research directions there. Finally the chapter ends with the conclusion.

BACKGROUND

The focus of this chapter is to compare five existing AC approaches in the application of data analysis on MSC differentiation. AC in fact lies at the overlap between classification and ARM, which solves the traditional classification problem based on ARM techniques with regard to rule generation and presentation. As mentioned above, AC has been selected as a suitable technique for MSC differentiation analysis. In this section, the authors concentrate scientifically and technically on the depiction of (*single-label multi-class*) classification, ARM and AC.

Classification

Classification is a traditional school in the field of data mining, as well as in machine learning. It is a typical form of "*data analysis that can be used to extract models describing important data classes*" (Han & Kamber, 2006). Specifically, classification

aims to assign predefined data categories/classes to "unseen" data instances, based on the study of a given set of training data examples — data instances associating with data category labels. Early studies of (data) classification can be dated back to the early 1960s, see for instance (Maron, 1961) with regard to such textual data. The process of classification consists of two steps: (1) "*a classifier is built describing a predetermined set of data classes or concepts*" — "*this is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or 'learning from' the training set*" (Han & Kamber, 2006); and (2) the classifier model is used for classifying "unseen" data samples into predefined classes as given in the training set — this is the classification step (or test phase). In step (2), the measure of accuracy has been widely used to evaluate the performance of classification, especially as presented in this chapter when dealing with the *single-label multi-class* classification task. Agrawal, Gadbole, Punjani and Roy (2007) confirm that "*in a classification problem, the classification system is trained on the training data and effectiveness is measured by accuracy on test data"*, which is the fraction of correctly predicted instance-class mappings.

Broadly speaking, mechanisms on which classification algorithms have been based can be separated into two "families": *direct classification* — classification without rule generation; and *rule based classification* — classification with rule generation (and presentation). The "family" of *direct classification* focuses on directly classifying "unseen" data instances into predefined categories, but has no concern for presenting to the end user why and how the classification predictions have been made. Since this group of mechanisms only aims to show that machines can learn and make correct classification decisions, such (Classification) approaches were proposed under the machine learning heading. In *direct classification*, typical mechanisms include: naive Bayes (Lowd & Domingos, 2005), SVM (Boser, Guyon, & Vapnik, 1992), genetic algorithm (Freitas, 2002; Yang, Widyantoro, Ioerger, & Yen, 2001), neural networks (Han & Kamber, 2006), etc.

The "family" of *rule based classification* mines and generates human readable Classification Rules (CRs) from a given class-database D_C, with the objective of building a classifier to categorize "unseen" data records. Such mechanisms in this "family" were proposed under the data mining heading. Generally, D_C is described by a relational database table that includes a class attribute — whose values are a set of predefined class labels $C = \{c_1, c_2, \ldots, c_{|C|-1}, c_{|C|}\}$. The *two-step* process of *rule based classification* can be described formally as (1) CRs are generated from a set of training data instances $D_R \subset D_C$; and (2) "unseen" instances in a test dataset $D_E \subset D_C$ are assigned into predefined class groups. A D_C is established as $D_R \cup D_E$, where $D_R \cap D_E = \varnothing$. Both D_R and D_E share the same database attributes except the class attribute. By convention the last attribute in each D_R record usually indicates the predefined class of this record, noted as the class attribute, while the class attribute is missing in D_E. Typical mechanisms in *rule based classification* include: Decision Trees (Quinlan, 1993), *k*-NN (James, 1985), *FOIL* (Quinlan & Cameron-Jones, 1993), RIPPER (Cohen, 1995), Association Rules (Liu, et al., 1998; Wang, Xin, & Coenen, 2008), etc.

Association Rule Mining

Association Rule Mining (Cody et al. 2000), first introduced by Agrawal et al. (1993), aims to extract a set of Association Rules (ARs) from a given transactional database D_T. Association Rule describes an implicative co-occurring relationship between two sets of *binary*-valued (i.e. ABSENCE or APPEARANCE, 0 or 1) transactional database attributes (items), expressed in the form of an "antecedent ⇒ consequent" rule. As indicated by Cornelis et al. (2006), the concept of mining ARs can be dated back to work in the 1960's (Hajek, Havel, & Chytil, 1966).

In a more general form, ARM can be defined as follows. Let $I = \{a_1, a_2, \ldots, a_{n-1}, a_n\}$ be a set of items, and $\mathcal{T} = \{T_1, T_2, \ldots, T_{m-1}, T_m\}$ be a set of transactions (data records), a transactional database D_T is described by $\mathcal{T}$, where each $T_j \in \mathcal{T}$ comprises a set of items $I' \subseteq I$. In ARM, two threshold values are usually used to determine the significance of an AR:

- ***Support*:** A set of items S is called an itemset. The *support* of S is the proportion of transactions T in $\mathcal{T}$ for which $S \subseteq T$. If the *support* of S exceeds a user-supplied *support* threshold σ, S is defined as a *frequent itemset*.
- ***Confidence*:** *Confidence* represents how "strongly" an itemset X implies another itemset Y, where $X, Y \subseteq I$ and $X \cap Y = \varnothing$. A *confidence* threshold α, supplied by a user, is used to distinguish high confidence ARs from low confidence ARs.

An AR $X \Rightarrow Y$ is said to be *valid* when the *support* for the co-occurrence of X and Y exceeds σ, and the *confidence* of the AR exceeds α. The computation of *support* is:

$$support(X \cup Y) = count(X \cup Y) / |\mathcal{T}|,$$

where $count(X \cup Y)$ is the number of transactions containing the set $X \cup Y$ in $\mathcal{T}$, and $|\mathcal{T}|$ is the *size* function (*cardinality*) of the set $\mathcal{T}$. The computation of *confidence* is:

$$confidence(X \Rightarrow Y) = support(X \cup Y) / support(X).$$

Informally, "$X \Rightarrow Y$" can be interpreted as: if X is found in a transaction, it is likely that Y also will be found.

In general, ARM involves a search for all *valid* rules. The most computationally difficult part of this is the identification of *frequent itemsets*. Since its introduction in 1994, the *apriori* algorithm developed by Agrawal and Srikant (1994) has been the basis of many subsequent ARM algorithms. In Agrawal and Srikant (1994) it was observed that ARs can be straightforwardly generated from a set of *frequent itemsets*. Thus, efficiently and effectively mining *frequent itemsets* from data is the key to ARM. The *apriori* algorithm iteratively identifies *frequent itemsets* in data by employing the "closure property" of itemsets in the generation of candidate itemsets, where a candidate (possibly frequent) itemset is confirmed as frequent only when all its subsets are identified as frequent in the previous pass. The "closure property" of itemsets can be described as follows: if an itemset is frequent then all its subsets will also be frequent; conversely if an itemset is infrequent then all its supersets will also be infrequent.

With regards to the history of ARM investigation, many algorithms have been introduced that mine ARs from identified *frequent itemsets*. These algorithms can be further grouped into different "families", such as Pure-*apriori* like, Semi-*apriori* like, Set Enumeration Tree like, etc.

- **Pure-*apriori* like** were *frequent itemsets* are generated based on the generate-prune level by level iteration that was first promulgated in the *apriori* algorithm. In this "family" archetypal algorithms include: *apriori*, *apriori*-Tid and *apriori*-Hybrid (Agrawal & Srikant, 1994), Partition (Savasere, Omiecinski, & Navethe, 1995), Sampling (Toivonen, 1996), DIC (Brin, Motwani, Ullman, & Tsur, 1997), CARMA (Hidber, 1999), etc.
- **Semi-*apriori* like** were *frequent itemsets* are generated by enumerating candidate itemsets but do not apply the *apriori* generate-prune iterative approach founded in (1) the join procedure, and (2) the prune procedure that employs the "closure property" of itemsets. In this "family" typical algorithms include: AIS (Agrawal et al., 1993), SETM (Houtsma & Swami, 1995),

OCD (Mannila, Toivonen, & Verkamo, 1994), etc.

- **Set Enumeration Tree like** were *frequent itemsets* are generated through constructing a *set enumeration tree* structure (Rymon, 1992) from D_T, which avoids the need to enumerate a large number of candidate itemsets. In this "family" a number of approaches can be further divided into two main streams: *apriori-TFP*[1] based (Coenen, Goulbourne, & Leng, 2001; Coenen & Leng, 2002; Coenen, Leng, & Ahmed, 2004; Coenen, Leng, & Goulbourne, 2004), and *FP-tree* based (El-Hajj & Zaiane, 2003; Han, Pei, & Yin, 2000; Liu, Pan, Wang, & Han, 2002).

Associative Classification

An overlap between ARM and *rule based classification* is AC (Associative Classification) or CARM (Classification Association Rule Mining), which strategically solves the traditional classification problem by applying ARM techniques. The idea of AC, first introduced in (Ali, Manganaris, & Srikant, 1997), aims to extract a set of Classification Association Rules (CARs) from a class-transactional database $D_{C\text{-}T}$. Let D_T be a transactional database, and $C = \{c_1, c_2, \ldots, c_{|C|-1}, c_{|C|}\}$ be a set of predefined class labels, $D_{C\text{-}T}$ is described by $D_T \times C$. $D_{C\text{-}T}$ can also be defined as a special class-database D_C, where all database attributes and the class attribute are valued in a *binary* manner — "*Boolean attributes can be considered a special case of categorical attributes*" (Srikant & Agrawal, 1996). A CAR is a special AR that describes an implicative co-occurring relationship between a set of *binary*-valued data attributes and a predefined class, expressed in the form of an "$X \Rightarrow c_i$" rule, where X is an itemset found in D_T (as "$D_{C\text{-}T} - C$") and c_i is a predefined class in C.

AC offers the following advantages with respect to the classification techniques mentioned above (Antonie & Zaiane, 2002; Yoon & Lee, 2005):

- The approach is efficient during both the training and categorization phases, especially when handling a large volume of data.
- The classifier built in this approach can be read, understood and modified by humans.

Furthermore, AC is relatively insensitive to noise data. AC builds a classifier by extracting a set of CARs from a given set of training instances. Possible CARs are determined by a large enough *support* and a large enough *confidence*. Usually, rules derived from noise in the data will fail to reach these thresholds and will be discarded.

In comparison, classification approaches other than AC, i.e. naive Bayes, SVM, genetic algorithm, neural networks, etc. do not present the classification in a human readable fashion, so that users do not see why the (Classification) predictions have been made by computers. While rules generated by decision tree classifier, RIPPER classifier, etc. can be read and understood by humans, however (Yin & Han, 2003) report that in many cases AC offers higher classification accuracy than other *rule based classification* approaches.

For these reasons it was decided to use an AC approach to address the prediction of mammalian MSC differentiation. One of the existing AC frameworks is the CMAR (Classification based on Multiple Association Rules) algorithm (Li et al., 2001). CMAR generates CARs (from a given set of training instances) through an *FP–tree* (Han, Pei, & Yin, 2000) based approach. Experimental results using this algorithm show that it could achieve high classification accuracy for a range of data sets (Li et al., 2001). Other alternative AC techniques are CBA, PRM, CPAR, TFPC, etc.

FIVE ASSOCIATIVE CLASSIFICATION APPROACHES

Classification Based on Associations

The Classification Based on Associations (CBA) algorithm (Liu et al., 1998) exemplifies the "two stage" approach (as opposite to the "integrated" approach), and was one of the first to make use of a general ARM algorithm for "stage 1". CBA uses a version of the well-known *apriori* algorithm (Agrawal & Srikant, 1994), using user-supplied *support* and *confidence* thresholds, to generate CARs which are then prioritized as follows (e.g. given two rules r_A and r_B):

- r_A has priority over r_B if the *confidence* value of r_A is greater than the *confidence* value of r_B.
- r_A has priority over r_B if the *confidence* values of r_A and r_B are equal, but the *support* value of r_A is greater than the *support* value of r_B.
- r_A has priority over r_B if the *confidence* values of r_A and r_B are equal, the *support* values of r_A and r_B are equal, but the rule-antecedent *size* (the number of items) of r_A is less than the rule-antecedent *size* of r_B.

In "stage 2", the ordered set of CARs is then pruned as follows:

- For each data record *d* in the training set, find the first CAR (the one with the highest precedence) that correctly classifies the record (the *cor-CAR*), and the first CAR that wrongly classifies the record (the *wro-CAR*).
- For each data record where the *cor-CAR* has higher precedence than the *wro-CAR*, such CARs are included in the classifier.
- For all data records where the *cor-CAR* does not have higher precedence than the *wro-CAR*, alternative CARs with lower precedence must be considered and added to the classifier.

CARs are added to the classifier according to their precedence. On completion the lower precedence CARs are examined and a default rule selected to replace these low precedence CARs. CBA illustrates the general performance drawback of "two stage" algorithms — the cost of the pruning stage is a product of the size of the data set and the number of candidate CARs, both of which may in some cases be large. It is clear, also, that the choice of *support* and *confidence* thresholds will strongly influence the operation of CBA. The ordering strategy, noted as *Confidence-Support-Antecedent* (CSA), seems to work well on some data sets.

Classification Based on Multiple Association Rules

The Classification based on Multiple Association Rules (CMAR) algorithm (Li et al., 2001) has a similar general structure to CBA, and uses the same rule prioritization approach as that employed in CBA. CMAR differs in the method used in "stage 1" to generate candidate CARs, which makes use of the *FP-tree* data structure coupled with the *FP-growth* algorithm (Han et al., 2000); this makes it more computationally efficient than CBA. Like CBA, CMAR tends to generate a large number of candidate CARs. The set of CARs is pruned by removing all rules with a χ squared value below a user-defined threshold and all rules where a more general rule with higher precedence exists. Finally, a database coverage procedure is used to produce the final set of CARs. This stage is similar to that of CBA, but whereas CBA finds only one CAR to cover each case, CMAR uses a coverage threshold parameter to generate a large number of CARs. When classifying an "unseen" data record, CMAR groups CARs that satisfy the record according to their class and determines the

combined effect of the CARs in each group using a Weighted χ Squared (WCS) measure.

Predictive Rule Mining

Predictive Rule Mining (PRM) (Yin & Han, 2003), as an extension of the *FOIL* algorithm (Quinlan & Cameron-Jones, 1993), is a time-efficient algorithm based on the *greedy* paradigm in which rules to distinguish positive examples from negative ones are iteratively learnt. PRM repeatedly searches for the current "best" rule and decreases the weights of the positive examples when those positive examples are correctly covered by this selected "best" rule until all the positive examples in the (training) data set are covered. Note that (in comparison) the positive examples will be removed if such examples are covered by any selected "best" rule during each iteration of the rule selection in traditional *FOIL*. By performing such an approach, PRM can produce more rules than *FOIL* and each positive example is usually covered more than once. Consequently, it leads higher classification accuracy than *FOIL*. Similarly as the methodologies used in *FOIL*, a crucial function *gain* (*p*) is used to measure the information gained from adding the literal *p* to the current rule *r* during selection of literals, e.g. the number of bits saved in representing all the positive examples by adding *p* to *r*. In order to achieve a better efficiency than *FOIL*, PRM employs the standard approach (Gehrke, Ramakrishnan, & Ganti, 1998) based on a new data structure called *PNArray* to retail the computational burden on evaluation of every literal during searching stages for the one with the highest gain in *FOIL*. For *multi-class* (Classification) problems, PRM follows the standard framework from *FOIL*: for each class, its examples are used as positive examples and those of other classes as negative ones, and the rules for all classes are merged together to form the classifier (rule set).

Classification Based on Predictive Association Rules

Classification based on Predictive Association Rules (CPAR) (Yin & Han, 2003) inherits the basic idea of traditional *FOIL* in rule generation and integrates the features of associative classification of PRM. When selecting literals during the rule building process, PRM selects only one "best" literal in each iteration and ignores all the others. In fact, there are usually many rules with similar accuracy based on the remaining dataset in each iteration. The "best" rule among them in the remaining dataset may not be the "best" rule in the whole (training) dataset. This strategy may therefore lead to PRM missing some very important rules. Instead of ignoring all literals except the "best" one, CPAR keeps all close-to-the-best literals in each iteration during the rule building process. By performing such an approach, CPAR can select more than one literal at the same time and build several rules simultaneously. In comparison with PRM, CPAR has the following advantages: (1) CPAR generates a much smaller set with high-quality predictive rules directly from the given dataset; (2) to avoid producing redundant rules, CPAR generates each rule by taking into account the set of "already-generated" rules; and (3) when predicting the class label for a given example, CPAR uses the *best k rules* on which this example satisfies.

Total From Partial Classification

Several of the above AC methods apply coverage analysis to prune data instances/cases and reduce the number of rules generated in the training phase. It can be demonstrated that coverage analysis, especially when applied to a large $D_{C\text{-}T}$ comprising many items and multiple transactions, includes a significant computational overhead. This is the motivation behind development of an algorithm that directly builds an acceptably accurate classifier without coverage analysis. The Total From

Partial Classification (TFPC) algorithm, proposed by Coenen et al. (2005), is directed at this aim. Coenen and Leng (2007) argue that the principal advantage offered by TFPC is that "*it is extremely efficient (because it dispenses with the need for coverage analysis)*".

TFPC is derived from the *apriori-TFP* ARM approach (Coenen, 2004; Coenen & Leng, 2004). It employs the same (*set enumeration tree*) structures and (mining) procedures as used in *apriori-TFP* to the task of identifying CARs in $D_{C\text{-}T}$. For this purpose, predefined class labels in $D_{C\text{-}T}$ are considered as items, and set at the end of the item list (ordered in a descending manner based on the item frequency).

In its rule generation process, TFPC adopts the heuristic: "*if we can identify a rule $X \Rightarrow c$ which meets the required support and confidence thresholds, then it is not necessary to look for other rules whose antecedent is a superset of X and whose consequent is c*" (Coenen et al., 2005). The advantages of employing this heuristic can be listed as follows.

- It "*reduces the number of candidate rules to be considered*" thus "*significantly improving the speed of the rule-generation algorithm*" (Coenen & Leng, 2007).
- It reduces the number of final CARs to be generated, so that "*this 'on-the-fly' pruning replaces the expensive pruning step that other algorithms perform by coverage analysis*" (Coenen & Leng, 2007).
- It reduces the risk of *over-fitting* — i.e. the risk of producing a set of CARs that perform well on the training dataset but do not generalize well to the test dataset.

The classifier built by TFPC is finally represented as a list of CARs in a CSA rule ordering fashion. When classifying "unseen" cases TFPC typically uses the *best first rule* approach.

DATA PREPARATION

Parameter Selection from the Online MSC Database

In order to integrate mammalian MSC differentiation data, an online database[2] containing over 500 parameters that are believed to influence the MSC differentiation has been built in the previous studies (Wang et al., 2009). All the data in this online database have been published in the literature and marked with their respect references.

The current size of this database is 501 records, covering four types of MSC differentiation fates as predefined classes, which are osteogenesis, chondrogenesis, adipogenesis and proliferation without differentiation. The total number of parameters in this database is up to 500, including those which are believed to be most significant, such as donor species, *in vitro* vs. *in vivo* culture, culture medium, supplements and growth factors, culture dimension (monolayer vs. 3D culture), substrate (for monolayer culture) vs. scaffold (for 3D culture), those which are believed to be potentially important, such as age of donor, cell passage number, cell seeding density, incubation duration, those which usually act as supplementary comments, such as donor gender, MSC harvest place, and those representing cell behaviors as experimental results, including MSC differentiation fates, population doubling time, expression of cell markers, gene profiles, expansion fold of cell number, etc.

Among all the parameters in the database, those which are believed to be the most essential ones were abstracted and considered in this study. Table 1 shows all the parameters used for prediction in the current stage of this study. Consequently, the number of parameters in the abstracted database was reduced from 500 to 105.

Table 1. The abstracted database

Parameter groups	Significance/Description
Donor species	MSCs from different species of mammal in the same culture condition may lead to different results. The current database covers five different donor species.
Culture medium	The most essential of environment conditions where MSCs grow, proliferate and differentiate. A different culture medium has different effect on cells. The current database covers 16 types of culture media.
Supplements and growth factors	Chemicals that maintain MSC differentiation potential or influence their differentiation fates. The functions and effects of growth factors on MSCs vary from one to another, leading to different experimental results. The current database covers 64 types of supplements.
Culture dimension (2D vs. 3D) MSC differentiation sometimes differs significantly from monolayer to 3D culture, even under the same culture medium and supplements. This is one parameter with two possible values. Substrate (for 2D) /scaffold (for 3D)	Influences cell viability. A chemically modified substrate can even change MSCs' differentiation fate. The current database covers 10 types of substrates and 5 types of scaffolds.
Differentiation fate	To what lineage MSCs are committed to after differentiation. It is the most significant result after cell culture. Used to define the classes in the database; the objective of this study is to predict it. The current database covers four types of differentiation fates, as four classes.

Data Normalization and Cleaning

After parameter selection, the database was discretised and normalized using the LUCS-KDD Discretised Normalized (DN) software[3], so that data was made available in *binary* form and suitable for use by AC applications. In this study, the discretisation and normalization processes result in a data file with its number of attributes increased to 183.

This discretised and normalized data file contains noisy data, generally caused by the absence of culture condition parameters such as culture media, supplements & growth factors[4], etc. For example, if the insulin growth factor is absent in a record, this record will have an attribute representing "absence of insulin" after the discretisation and normalization process. This kind of attributes does not provide any useful information while increasing the complexity of the data file. Thus, all the attributes with a value of "absence" were eliminated, with the resulting data file referred as the preliminary data file.

Data Pre-Processing

The preliminary data file was not directly used as input data file to the five AC approaches because it contains some overlapping attributes. For example, some records in the preliminary data file contain an attribute for the presence of "ITS-plus", because in some experiments "ITS-plus" is used as supplement to the culture medium ("ITS-plus" is a combination of 6.25 g/ml of bovine insulin, 6.25 g/ml of transferrin, 6.25 g/ml of selenous acid, 5.33 g/ml of linoleic acid, and 1.25 mg/ml of bovine serum albumin39). In this case, the attribute for "ITS-plus" overlaps with the attribute for "insulin", and hence should be converted into one attribute for "insulin" plus four more attributes for the other four chemicals indicated above. On the other hand, some attributes in the data file are not useful. For example, it is known that the pres-

Table 2. Pruning of the attributes in the preliminary data file during data pre-processing

Attributes before pre-processing	Attributes after pre-processing	Ref.
antibiotic-antimycotic, penicillin, streptomycin, gentamicin	none	n/a
L-glutamine, glutamine	glutamine	n/a
platelet lysate	PDGF-αα, PDGF-ββ, TGF-β, VEGF, EGF	(Celotti, Colciago, Negri-Cesi, Pravettoni, Zaninetti, & Sacchi., 2006; O'Connell, Impeduglia, Hessler, Wang, Carroll, & Dardik, 2008)
ITS-plus/ITS+permixTX, ITS+1	insulin, transferrin, selenous acid, LA-BSA	(Johnstone, Hering, Caplan, Goldberg, & Yoo, 1998; Mackay, Beck, Murphy, Barry, Chichester, & Pittenger, 1998)
SITE (from sigma)	selenous acid, insulin, transferrin, ethanolamine	(Liu, Wu, & Hwang, 2007)
ascorbic acid, scorbate-2-phosphate/ascorbic acid-2-phosphate	"ascorbic acid (-2-phosphate)"	n/a
IBMX, 8-MM-IBMX	"IBMX or 8-MM-IBMX"	n/a
TGF-β1, TGF-β3	TGF-β	n/a

ence of the supplement "antibiotic-antimycotic" is to prevent contamination and has no influence on MSC differentiation. Attributes concerned with this type of supplements should hence be eliminated from the preliminary data file. As a result of pruning the attributes according to Table 2, the preliminary data file became the input data file to the five AC approaches, with 95 attributes in total. This step is referred as data pre-processing, after which the five AC approaches were applied to segments of the input data file with a Ten-fold Cross Validation (TCV) accuracy setting (90% training set, 10% test set) (Schaffer, 1993), with the results shown in the next section.

EXPERIMENTS

In this work, comparison of five different AC approaches on the performance in MSC data analysis has been focused. The five AC approaches for comparison are CBA, CMAR, PRM, CPAR and TFPC. Experiments were run on a 2.00 GHz Intel(R) Core(TM) 2 CPU with 2.00 GB of RAM running under Windows Command Processor. The TCV evaluation undertaken used a *confidence* threshold value (α) of 50% and a *support* threshold value (σ) of 1% for CBA, CMAR and TFPC with the intension to avoid *over-fitting* (Coenen & Leng, 2004; Coenen et al., 2005; Li et al., 2001; Wang, Xin, & Coenen, 2007), while for PRM and CPAR there is no such notion of *support* and *confidence*.

Performance Comparison

After the five AC approaches were applied to the input data file with TCV, each of them showed different performance, in terms of (1) classification accuracy in each TCV fold and the average accuracy, (2) number of CARs generated in each TCV fold and the average number of CARs generated, (3) maximum number of attributes in CAR antecedents, and (4) generation time after which the classification was accomplished.

The average accuracy, average number of CARs and generation time for each AC approach are shown in Table 3. Among all, CBA gave the highest average accuracy of 94.8%, while the lowest accuracy of 67.6% was obtained from CPAR. In terms of average number of CARs, CMAR showed a preeminent result of 290.7, while no other AC approach gave a number higher than 81.5. The sort ascending order of generation time for the five AC approaches is PRM, CPAR, CBA,

Table 3. Performance of the five AC approaches

	CBA	CMAR	PRM	CPAR	TFPC
Average accuracy (%)	94.8	90.42	69.0	67.6	90.6
Average num of CRs	58.8	290.7	34.6	38.0	81.5
Max. num of attributes in antecedents	4	6	4	4	3
Generation time (seconds)	2.58	57.61	0.33	0.55	2.73

TFPC and CMAR; however, CMAR has the largest maximum number of attributes in antecedents of 6, remarkably more than the other AC approaches.

Rule Comparison

For every independent AC approach, a number of rules were generated in each TCV fold, based on which AC makes classification prediction and evaluates the accuracy. These rules are important to this study because they may contain useful and valuable information (or knowledge) on stem cell differentiation. For the fairness of the comparison, the authors simply chose to compare the rules in the respect fold No.10 of each tested AC approach. Three most interesting rules in each AC approach were selected manually and listed in a descending order with the respect interpretation. The reader is reminded that for the rules from CBA, CMAR and TFPC, their *confidence* values were shown in square brackets, while the rules from PRM and CPAR do not have *confidence* values due to their algorithms. The evaluation to the rules in terms of their significance was elucidated in the next section, based on *a priori* knowledge.

- In CBA, the following rules were believed to be most interesting:
  1. **Rule # CBA20:** {FBS + ascorbic acid + dexamethasone + TCP} ⇒ {osteo} [100.0%], which can be interpreted as: in the presence of FBS (Fetal Bovine Serum), ascorbic acid and dexamethasone, MSCs will undergo osteogenesis on the substrate of TCP (Tissue Culture Plastic).
  2. **Rule # CBA40:** {transferrin + dexamethasone + TGF-β + TCP} ⇒ {chondro} [94.73%], interpreted as: with the help of transferrin, dexamethasone, TGF-β in the culture medium, MSCs is most likely to differentiate into cartilage on TCP substrate.
  3. **Rule # CBA13:** {β-glycerophosphate + BMP-2} ⇒ {osteo} [100.0%], meaning that the combination of β-glycerophosphate and BMP-2 always stimuli MSCs to become bone cells.
- In CMAR, the selected rules are listed as follows:
  1. **Rule#CMAR89:** {FBS + ascorbic acid + dexamethasone + β-glycerophosphate + 2D + TCP} ⇒ {osteo} [100.0%], meaning that MSC cultured on plastic substrate in monolayer culture will be induced into osteogenesis if supplemented with FBS, dexamethasone, β-glycerophosphate and ascorbic acid.
  2. **Rule#CMAR154:** {human + ascorbic acid + insulin + TGF-β} ⇒ {chondro} [96.42%], meaning that human MSC is most likely to undergo chondrogenic differentiation under the stimuli of the combined treatment with insulin and TGF-β together with ascorbic acid (or ascorbic acid-2-phosphate).
  3. **Rule # CMAR127:** {human + FBS + dexamethasone + insulin + 2D + TPC} ⇒ {adipo} [100%], suggesting that the culture conditions above is supportive for human MSC adipogenesis.

- In PRM, the selected rules are:
  1. **Rule # PRM1:** {DMEM + dexamethasone + β-glycerophosphate} ⇒ {osteo}, meaning that in the presence of the culture medium and supplements as above, MSCs will undergo osteogenesis.
  2. **Rule # PRM14:** {human + DMEM-HG + FBS + TCP} ⇒ {prolife}, meaning that DMEM-HG and FBS only helps human MSCs proliferate, without inducing them to any type of differentiation.
  3. **Rule # PRM9:** {TGF-β + proline} ⇒ {chondro}, meaning that TGF-β and proline may together promote chondrogenesis.
- In CPAR, the rules were exactly the same as those in PRM, except that the following three rules were found not to exist in PRM:
  1. **Rule # CPAR27:** {UltroserG serum substitute} ⇒ {prolife}, suggesting that the UltroserG serum substitute does not induce MSC differentiation.
  2. **Rule # CPAR33:** {DMEM-F12} ⇒ {osteo}, suggesting that DMEM-F12 may be biased on osteogenesis rather than other types of differentiation.
  3. **Rule # CPAR38:** {dexamethasone} ⇒ {chondro}, suggesting that in the current database, dexamethasone appears more frequently in chondrogenesis than other differentiation types.
- In TFPC, the selected rules are:
  1. **Rule # TFPC66:** {FBS + dexamethasone + insulin} ⇒ {adipo} [54.54%], meaning that in many cases the combination of FBS, dexamethasone and insulin can differentiate MSCs into fat cells, but not always.
  2. **Rule # TFPC46:** {DMEM-HG + ascorbic acid} ⇒ {chondro} [69.44%], meaning that ascorbic acid in the culture medium of DMEM-HG can promote chondrogenesis.
  3. **Rule#TFPC1:** {proline} ⇒ {chondro} [100.0%], suggest that proline may play a role in chondrogenesis.

As listed above, all the AC approaches can abstract classification rules containing information on MSC differentiation. However, the quality of these rules, in terms of the extent to which they correlate with a priori knowledge and how well the rules were integrated, has to be evaluated manually, as elucidated in the next section.

FUTURE RESEARCH DIRECTIONS

Five AC approaches have been used in this study with the aim of comparing their performance on prediction of MSC differentiation by classification, and abstraction of hidden rules from currently available experimental data.

Results from all the five AC approaches have been derived in terms of both computational performance (i.e. classification accuracy, number of CARs generated, maximum number of attributes in CAR antecedents, and time efficiency) and CARs abstracted from MSC data. From Table 3 it can be seen that all the classifiers gave accuracy higher than 90% except PRM and CPAR. However, CMAR showed a preeminent result of 290.7 on average number of CARs, with the largest maximum number of attributes in CAR antecedents of 6, remarkably higher than the other classifiers. Despite that the generation time of CMAR is the longest, the best AC approaches suggested in this study relies on the balance of the four types of performance and the quality of the mined CARs.

For CBA, all the three selected rules are consistent with observations in lab; however, the Rule # CBA20 is obviously not as good as Rule # CMAR89, because the former one is a subset of the latter one. In fact, due to the limited size

of the CBA rules, CMAR excelled CBA in the similar cases for some other rules as well (data not shown). For PRM and CPAR, these two algorithms gave exactly the same rules, except for Rules # CPAR27, 33 and 38. After being analyzed, none of these three rules showed valuable information. For example, dexamethasone was known to participate also in osteogenesis and adipogenesis, whereas Rule # CPAR38 claims it to be only beneficial to chondrogenesis. Thus, CPAR is close to PRM on rule quality.

Among all the five classifiers, TFPC gave the most limited average length of rules, with the maximum number of attributes in antecedents of three. This results in the problem that few of the TFPC rules were well integrated, although their *confidence* values were relative high. In fact, over half of the TFPC rules have only one attribute in their respect antecedent, which makes it extremely difficult to provide useful biological information. In contrast to TFPC, CMAR generates CARs in a most integrated manner. For example, Rule # CMAR89 gives a more integrated abstraction for osteogenesis than Rule # CBA20. Rule # CMAR154 is believed to be better organized than Rule # TFPC46, as the former one contains more information. In many more cases, CMAR also exceeded the other four AC approaches on the quality of rules.

From the analysis above, an overview of rule quality for each AC approach can be derived, which is that CMAR performs best in generating rules with integrated information. Although PRM and CPAR cost much less generation time, CBA provides the highest accuracy and TFPC has a good balance in time efficiency, accuracy and number of CARs, CMAR is suggested as the most suitable classifier to this study due to its excelling rule quality and satisfactory accuracy. However, CMAR has the same problem with the other classifiers, which is that a number of rules do not make scientific sense. For example, for Rule # CMAR204: {goat + 2D} ⇒ {osteo} [88.88%], it is obvious that only "Goat MSC" and "monolayer culture" are not enough to induce osteogenesis. Similarly, for Rule # CMAR274: {FBS + 3D} ⇒ {osteo} [72.41%], according to authors knowledge, FBS is not specifically for promoting osteogenesis but for maintaining cell survival without promoting effect towards any differentiation, independently of the fact that the culture is monolayer or 3D. In fact, all the five tested AC approaches have some rules without scientific sense. As a result, all the rules have to be reviewed by human beings for the rule quality. The generation of non-scientific rules is due to the size of the database and the sample properties of the data. Based on this reason, a conclusion can be made that if the MSC database is expanded in the future, the non-scientific rules could be pruned and more rules with scientific sense could be identified.

CONCLUSION

In this study, MSC data from an online database were processed and analyzed by five different AC approaches in order to compare their performance with respect to several aspects. Due to the capacity of AC, which is to harmonize the vast amount of experimental data and produce simple but useful rules, it is recommended as a suitable tool for this study. After the comparison between the five AC approaches, CMAR is suggested to be the most suitable approach for this study, and possibly also suitable to other similar studies such as the tissue engineering related data analysis. Due to the limited experimental data input at this stage, most of the identified rules are known by stem cell researchers. However, it will be possible to mine completely original rules if the size and contents of the MSC database are expanded in the future.

ACKNOWLEDGMENT

The authors would like to thank Professor Jian Lu from the School of Physics & Astronomy at the University of Manchester, and the following colleagues from the Department of Engineering Science at the University of Oxford for their valuable suggestions to this study: Paul Raju, Nuala Trainor, Dr. Cathy Ye, Dr. Xia Xu, Dr. Shengda Zhou, Dr. Renchen Liu, Professor James Triffitt, Clarence Yapp, Yang Liu and Zhiqiang Zhao.

The authors would also like to thank Professor Paul Leng from the Department of Computer Science at the University of Liverpool, Dr. Jiongyu Li and Fan Li from the Information Management Center in the China Minsheng Banking Corp. Ltd., and Zhijie Jia from the Beijing Friendship Hotel for their support with respect to the work described here.

REFERENCES

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rule between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* (pp. 207-216). Washington D.C.: ACM Press.

Agrawal, R., & Srikant, R. (1994). Fast Algorithm for Mining Association Rules. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB-94)* (pp. 487-499). Santiago de Chile, Chile: Morgan Kaufmann Publishers.

Agrawal, S., Godbole, S., Punjani, D., & Roy, S. (2007). How much noise is too much: A study in Automatic Text Classification. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM-07)* (pp. 3-12). Omaha, NE, USA: IEEE Computer Society.

Ali, K., Manganaris, S., & Srikant, R. (1997). Partial Classification using Association Rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (pp. 115-118). Newport Beach, CA, USA: AAAI Press.

Antonie, M. L., & Zaiane, O. R. (2002). Text Document Categorization by Term Association. In *Proceedings of the 2002 IEEE International Conference on Data Mining* (pp. 19-26). Maebashi City, Japan: IEEE Computer Society.

Battula, V. L., Bareiss, P. M., Treml, S., Conrad, S., Albert, I., & Hojak, S. (2007). Human Placenta and Bone Marrow derived MSC cultured in Serum-free, b-FGF-containing medium express cell surface frizzled-9 and SSEA-4 and give rise to multilineage differentiation. *Differentiation*, *75*(4), 279–291. doi:10.1111/j.1432-0436.2006.00139.x

Beeres, S. L., Atsma, D. E., van der Laarse, A., Pijnappels, D. A., van Tuyn, J., & Fibbe, W. E. (2005). Human Adult Bone Marrow Mesenchymal Stem Cells repair experimental conduction block in rat Cardiomyocyte Cultures. *Journal of the American College of Cardiology*, *46*(10), 1943–1952. doi:10.1016/j.jacc.2005.07.055

Bianco, P., Riminucci, M., Gronthos, S., & Robey, P. G. (2001). Bone Marrow Stromal Stem Cells: Nature, Biology, and Potential applications. *Stem Cells (Dayton, Ohio)*, *19*(3), 180–192. doi:10.1634/stemcells.19-3-180

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th ACM Annual Workshop on Computational Learning Theory* (pp. 144-152). Pittsburgh, PA, USA: ACM Press.

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic Itemset counting and Implication Rules for Market Basket Data. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data* (pp. 255-264). Tucson, Arizona, USA: ACM Press.

Celotti, F., Colciago, A., Negri-Cesi, P., Pravettoni, A., Zaninetti, R., & Sacchi, M. C. (2006). Effect of Platelet-rich plasma on Migration and Proliferation of SaOS-2 Osteoblasts: Role of Platelet-derived Growth Factor and Transforming Growth Factor-beta. *Wound Repair and Regeneration*, *14*(2), 195–202. doi:10.1111/j.1743-6109.2006.00110.x

Cody, G. D., Boctor, N. Z., Filley, T. R., Hazen, R. M., Scott, J. H., & Sharma, A. (2000). Primordial Carbonylated Iron-Sulfur compounds and the synthesis of Pyruvate. *Science*, *289*(5483), 1337–1340. doi:10.1126/science.289.5483.1337

Coenen, F. (2004). The LUCS-KDD Apriori-T Association Rule Mining Algorithm. D*epartment of Computer Science, The University of Liverpool, UK*. Retrieved from http://www.cxc.liv.ac.uk /~frans/ KDD/ Software/ Apriori_T/ aprioriT.html.

Coenen, F., Goulbourne, G., & Leng, P. (2001). Computing Association Rules using partial totals. In *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 54-66). Freiburg, Germany: Springer-Verlag.

Coenen, F., & Leng, P. (2002). Finding Association rules with some very Frequent Attributes. In *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 99-111). Helsinki, Finland: Springer-Verlag.

Coenen, F., & Leng, P. (2004). An Evaluation of Approaches to Classification Rule Selection. In *Proceedings of the 4th IEEE International Conference on Data Mining* (pp. 359-362). Brighton, UK: IEEE Computer Society.

Coenen, F., & Leng, P. (2007). The Effect of Threshold Values on Association Rule based Classification Accuracy. *Journal of Data and Knowledge Engineering*, *60*(2), 345–360. doi:10.1016/j.datak.2006.02.005

Coenen, F., Leng, P., & Ahmed, S. (2004). Data Structures for Association Rule Mining: T-trees and P-trees. *IEEE Transactions on Data and Knowledge Engineering*, *16*(6), 774–778. doi:10.1109/TKDE.2004.8

Coenen, F., Leng, P., & Zhang, L. (2005). Threshold Tuning for improved Classification Association Rule Mining. In *Proceedings of the 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 216-225). Hanoi, Vietnam: Springer-Verlag.

Coenen, F. P., Leng, P., & Goulbourne, G. (2004). Tree Structures for Mining Association Rules. *Journal of Data Mining and Knowledge Discovery*, *8*(1), 25–51. doi:10.1023/B:DAMI.0000005257.93780.3b

Cohen, W. W. (1995). Fast Effective Rule Induction. In *Proceedings of the 12th International Conference on Machine Learning* (pp. 115-123). Tahoe City, CA, USA: Morgan Kaufmann Publishers.

Cornelis, C., Yan, P., Zhang, X., & Chen, G. (2006). Mining Positive and Negative Association Rules from Large Databases. *Proceedings of the 2006 IEEE International Conference on Cybernetics and Intelligent Systems* (pp. 613-618). Bangkok, Thailand: IEEE Computer Society.

Derubeis, A. R., & Cancedda, R. (2004). Bone Marrow Stromal Cells (BMSCs) in Bone Engineering: Limitations and Recent Advances. *Annals of Biomedical Engineering*, *32*(1), 160–165. doi:10.1023/B:ABME.0000007800.89194.95

El-Hajj, M., & Zaiane, O. R. (2003). Inverted Matrix: Efficient Discovery of Frequent Items in Large Datasets in the context of Interactive Mining. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 109-118). Washington, DC: ACM Press.

Freitas, A. A. (2002). *Data Mining and Knowledge Discovery with Evolutionary Algorithm*. Germany: Springer-Verlag Berlin Heidelberg.

Gehrke, J., Ramakrishnan, R., & Ganti, V. (1998). RainForest: A Framework for Fast Decision Tree Construction of Large Datasets. In *Proceedings of International Conference on Very Large Data Bases* (pp. 416-427). New York, USA.

Griffith, L. G., & Swartz, M. A. (2006). Capturing complex 3D Tissue physiology in vitro. *Nature Reviews. Molecular Cell Biology*, *7*(3), 211–224. doi:10.1038/nrm1858

Hajek, P., Havel, I., & Chytil, M. (1966). The GUHA Method of Automatic Hypotheses Determination. *Computing*, *1*, 293–308. doi:10.1007/BF02345483

Han, J., & Kamber, M. (2006). *Data mining: Concepts and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

Han, J., Pei, J., & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. In P*roceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (pp. 1-12). ACM Press, Dallas, TX, USA.

Hanada, K., Dennis, J. E., & Caplan, A. I. (1997). Stimulatory effects of basic Fibroblast Growth Factor and Bone Morphogenetic Protein-2 on Osteogenic differentiation of Rat Bone Marrow-derived Mesenchymal Stem Cells. *Journal of Bone and Mineral Research*, *12*(10), 1606–1614. doi:10.1359/jbmr.1997.12.10.1606

Haynesworth, S. E., Baber, M. A., & Caplan, A. I. (1996). Cytokine expression by Human Marrow-derived Mesenchymal Progenitor Cells in vitro: Effects of Dexamethasone and IL-1 Alpha. *Journal of Cellular Physiology*, *166*(3), 585–592. doi:10.1002/(SICI)1097-4652(199603)166:3<585::AID-JCP13>3.0.CO;2-6

Hidber, C. (1999). Online Association Rule Mining. In *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data* (pp. 145-156). Philadelphia, Pennsylvania, USA: ACM Press.

Houtsma, M., & Swami, A. (1995). Set-oriented Mining of Association Rules in Relational Databases. In *Proceedings of the 11th International Conference on Data Engineering* (pp. 25-33). Taipei, Taiwan: IEEE Computer Society.

James, M. (1985). *Classification Algorithm*. New York: Wiley-Interscience.

Johnstone, B., Hering, T. M., Caplan, A. I., Goldberg, V. M., & Yoo, J. U. (1998). In vitro Chondrogenesis of Bone Marrow-derived Mesenchymal Progenitor Cells. *Experimental Cell Research*, *238*(1), 265–272. doi:10.1006/excr.1997.3858

Krampera, M., Glennie, S., Dyson, J., Scott, D., Laylor, R., & Simpson, E. (2003). Bone Marrow Mesenchymal Stem Cells inhibit the response of naive and memory antigen-specific T cells to their Cognate Peptide. *Blood*, *101*(9), 3722–3729. doi:10.1182/blood-2002-07-2104

Kuznetsov, S. A., Friedenstein, A. J., & Robey, P. G. (1997). Factors required for Bone Marrow Stromal Fibroblast Colony Formation in vitro. *British Journal of Haematology*, *97*(3), 561–570. doi:10.1046/j.1365-2141.1997.902904.x

Lennon, D. P., Haynesworth, S. E., Young, R. G., Dennis, J. E., & Caplan, A. I. (1995). A Chemically defined medium supports in vitro proliferation and maintains the osteochondral potential of Rat Marrow-derived Mesenchymal Stem Cells. *Experimental Cell Research*, *219*(1), 211–222. doi:10.1006/excr.1995.1221

Li, W., Han, J., & Pei, J. (2001). CMAR: Accurate and Efficient Classification based on Multiple Class-Association Rules. In *Proceedings of the 2001 IEEE International Conference on Data Mining* (pp. 369-376). San Jose, CA: IEEE Computer Society.

Liu, B., Hsu, W., & Ma, Y. (1998). Integrating Classification and Association Rule Mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (pp. 80-86). New York City: AAAI Press.

Liu, C. H., Wu, M. L., & Hwang, S. M. (2007). Optimization of Serum free medium for Cord Blood Mesenchymal Stem Cells. *Biochemical Engineering Journal*, *33*(1), 1–9. doi:10.1016/j.bej.2006.08.005

Liu, J., Pan, Y., Wang, K., & Han, J. (2002). Mining Frequent Item Sets by Opportunistic Projection. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 229-238). Edmonton, Alberta, Canada: ACM Press.

Lowd, D., & Domingos, P. (2005). Naive Bayes Models for Probability Estimation. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 529-536). Bonn, Germany: ACM Press.

Mackay, A. M., Beck, S. C., Murphy, J. M., Barry, F. P., Chichester, C. O., & Pittenger, M. F. (1998). Chondrogenic differentiation of Cultured Human Mesenchymal Stem Cells from Marrow. *Tissue Engineering*, *4*(4), 415–428. doi:10.1089/ten.1998.4.415

Magaki, T., Kurisu, K., & Okazaki, T. (2005). Generation of Bone Marrow-derived Neural Cells in Serum-free Monolayer Culture. *Neuroscience Letters*, *384*(3), 282–287. doi:10.1016/j.neulet.2005.05.025

Mannila, H., Toivonen, H., & Verkamo, A. I. (1994). Efficient Algorithms for Discovering Association Rules. In *Proceedings of the 1994 AAAI Workshop on Knowledge Discovery in Databases* (pp. 181-192). Seattle, Washington, USA: AAAI Press.

Maron, M. E. (1961). Automatic Indexing: An Experimental Inquiry. [JACM]. *Journal of the ACM*, *8*(3), 404–417. doi:10.1145/321075.321084

Meuleman, N., Tondreau, T., Delforge, A., Dejeneffe, M., Massy, M., & Libertalis, M. (2006). Human Marrow Mesenchymal Stem Cell Culture: Serum-free Medium allows better expansion than Classical Apha-MEM medium. *European Journal of Haematology*, *76*(4), 309–316. doi:10.1111/j.1600-0609.2005.00611.x

Muller, I., Kordowich, S., Holzwarth, C., Spano, C., Isensee, G., & Staiber, A. (2006). Animal Serum-free culture conditions for Isolation and Expansion of Multipotent Mesenchymal Stromal Cells from Human BM. *Cytotherapy*, *8*(5), 437–444. doi:10.1080/14653240600920782

O'Connell, S. M., Impeduglia, T., Hessler, K., Wang, X. J., Carroll, R. J., & Dardik, H. (2008). Autologous Platelet-rich Fibrin Matrix as Cell Therapy in the Healing of Chronic Lower-extremity Ulcers. *Wound Repair and Regeneration*, *16*(6), 749–756. doi:10.1111/j.1524-475X.2008.00426.x

Pittenger, M. F., Mackay, A. M., Beck, S. C., Jaiswal, R. K., Douglas, R., & Mosca, J. D. (1999). Multilineage potential of Adult Human Mesenchymal Stem Cells. *Science*, *284*(5411), 143–147. doi:10.1126/science.284.5411.143

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann Publishers.

Quinlan, J. R., & Cameron-Jones, R. M. (1993). FOIL: A Midterm Report. In *Proceedings of the 1993 European Conference on Machine Learning (ECML-93)* (pp. 3-20). Vienna, Austria: Springer-Verlag.

Roelen, B. A., & Dijke, P. (2003). Controlling Mesenchymal Stem Cell differentiation by TGFBeta Family Members. *Journal of Orthopaedic Science*, *8*(5), 740–748. doi:10.1007/s00776-003-0702-2

Rymon, R. (1992). Search through Systematic Set Enumeration. In *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning* (pp. 539-550). Cambridge, MA, USA: Morgan Kaufmann Publishers.

Savasere, A., Omiecinski, E., & Navathe, S. (1995). An Efficient Algorithm for Mining Association Rules in Large Databases. In *Proceedings of the 21st International Conference on Very Large Data Bases* (pp. 432-444). Zurich, Switzerland: Morgan Kaufmann Publishers.

Schaffer, C. (1993). Selecting a Classification Method by Cross-Validation. *Machine Learning*, *13*(1), 135–143. doi:10.1007/BF00993106

Shidara, Y., Nakamura, A., & Kudo, M. (2007). CCIC: Consistent Common Itemsets Classifier. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining (MLDM-07)* (pp. 490-498). Leipzig, Germany: Springer-Verlag.

Srikant, R., & Agrawal, R. (1996). Mining Quantitative Association Rules in Large Relational Tables. In *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data* (pp. 1-12). Montreal, Quebec, Canada: ACM Press.

Thabtah, F., Cowling, P., & Peng, Y. (2005). The Impact of Rule Ranking on the Quality of Associative Classifiers. In *Proceedings of AI-2005, the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence (AI-05) - Research and Development in Intelligent Systems XXII* (pp. 277-287). Cambridge, UK: Springer-Verlag.

Toivonen, H. (1996). Sampling Large Databases for Association Rules. In *Proceedings of the 22nd International Conference on Very Large Data Bases* (pp. 134-145). Mumbai (Bombay), India: Morgan Kaufmann Publishers.

Tuan, R. S., Boland, G., & Tuli, R. (2003). Adult Mesenchymal Stem Cells and Cell-based Tissue Engineering. *Arthritis Research & Therapy*, *5*(1), 32–45. doi:10.1186/ar614

Wang, W., Wang, Y. J., Bañares-Alcántara, R., Coenen, F., & Cui, Z. (in press). Analysis of Mesenchymal Stem Cell differentiation in vitro using Classification Association Rule Mining. *Journal of Bioinformatics and Computational Biology*.

Wang, W., Wang, Y. J., Bañares-Alcántara, R., Cui, Z., & Coenen, F. (2009). Application of Classification Association Rule Mining for Mammalian Mesenchymal Stem Cell differentiation. In *Proceedings of the 9th Industrial Conference on Data Mining (ICDM-09) — Advances in Data Mining Applications and Theoretical Aspects* (pp. 51-61). Leipzig, Germany: Springer-Verlag Berlin Heidelberg.

Wang, Y. J., Xin, Q., & Coenen, F. (2007). A Novel Rule Ordering Approach in Classification Association Rule Mining. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining* (pp. 339-348). Leipzig, Germany: Springer-Verlag.

Wang, Y. J., Xin, Q., & Coenen, F. (2008). Hybrid rule ordering in Classification Association Rule Mining. *Transactions on Machine Learning and Data Mining*, *1*(1), 1–15.

Yang, L., Widyantoro, D. H., Ioerger, T., & Yen, J. (2001). An entropy-based adaptive Genetic Algorithm for learning Classification Rules. In *Proceedings of the 2001 Congress on Evolutionary Computation (CEC-01)* (pp. 790-796). Seoul, South Korea: IEEE Computer Society.

Yin, X., & Han, J. (2003). CPAR: Classification based on predictive Association Rules. In *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM-03)* (pp. 331-335). San Francisco, CA, USA: SIAM.

Yoon, Y., & Lee, G. G. (2005). Practical Application of Associative Classifier for Document Classification. In *Proceedings of the Second Asia Information Retrieval Symposium* (pp. 467-478). Jeju Island, Korea: Springer-Verlag.

Zhang, Y., Li, C., Jiang, X., Zhang, S., Wu, Y., & Liu, B. (2004). Human placenta-derived Mesenchymal Progenitor Cells support culture expansion of long-term culture-initiating cells from Cord blood CD34+ cells. *Experimental Hematology*, *32*(7), 657–664. doi:10.1016/j.exphem.2004.04.001

KEY TERMS AND DEFINITIONS

Association Rule (AR): A typical knowledge model in data mining, which describes an implicative co-occurring relationship between two non-overlapping sets of binary-valued transactional database attributes.

Association Rule Mining (ARM): A research field in data mining, which aims to extract association rules from a given transactional database.

Associative Classification (AC): An overlap between classification and association rule mining that solves the traditional classification problem by applying association rule mining techniques.

Cell differentiation: The process by which a less specialized cell becomes a more specialized cell type. For example, a multipotent MSC becomes an osteoblast (specialized in bone generation).

Classification Association Rule (CAR): A special association rule that describes an implicative co-occurring relationship between a set of binary-valued transactional database attributes and one or more predefined data categories.

Classification Rule (CR): A typical knowledge model in data mining, which describes an implicative relationship between data attributes and predefined data categories.

Classification: A research field in data mining, which aims to assign predefined data categories to "unseen" data instances, based on the study of a given set of training data examples associating with category labels.

Confidence: The support of an association rule in relation to the support of its antecedent.

Mesenchymal Stem Cells (MSCs): Multipotent stem cells that can differentiate into a variety of cell types. Cell types that MSCs have been shown to differentiate into include osteoblasts, chondrocytes, myocytes, adipocytes, endotheliums, etc.

Support: The overall frequency in a given transactional database where an association rule applies.

ENDNOTES

1. The *apriori-TFP* and its related softwares may be obtained from http://www.csc.liv.ac.uk/~frans/KDD/Software
2. The online MSC database can be visited from http://www.oxford-tissue-engineering.org/forum/plugin.php?identifier=publish&module=publish
3. LUCS-KDD DN software may be obtained from http://www.csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN/lucs-kdd_DN.html
4. For related information, please find from http://www.oxford-tissue-engineering.org/forum/table3.doc