Sperm Donor Selection using Nominal and Binary Variable Methods

Rajendra B. Patil
School of Computer Sciences
North Maharashtra University,
Jalgaon (MS), India

B.V. Pawar
School of Computer Sciences
North Maharashtra University,
Jalgaon (MS), India

Ajay S. Patil
School of Computer Sciences
North Maharashtra University,
Jalgaon (MS), India

ABSTRACT
Technological advances in medical sciences are motivating researchers across the world to contribute to the field of assisted reproductive techniques, commonly referred as ART. During ART treatment, in case of male infertility, couples are often advised conceiving through donor sperms. Such couples have a common fear that the offspring may not resemble them in physical appearance. Any notable deviation in the offspring's physical appearance from his or her parents may seriously affect the couple both socially and psychologically. It is often expected that the selected sperm donor profile for a couple should have physical characteristics similar to either of the partners, most preferably the male. The ART specialist and sperm banks find difficulties in selecting donor profile that suitably matches with the requirements of the recipient couple. In this paper, nominal and binary variable methods are applied to identify the donor profiles that match with the requirements of recipient couple. The results are analytically presented and tested under the supervision of experts. It is found that the outcomes are satisfactory compared to manual selection and as a result these techniques can be integrated in designing an expert tool to assist the ART specialist and sperm banks in selecting the best matching donor profiles for recipient couples.

General Terms
ART (Assisted Reproductive Technology), ICMR (Indian Council of Medical Research)

Keywords
Infertility, ART (assisted reproductive technology), nominal method, binary variable method, sperm bank, donor profile matching

1. INTRODUCTION
There are over millions of infertile couples in the world and around 3-4 lakhs couples in India report for consulting fertility treatment every year [1][2][3]. In Indian society, the infertile couples face high stress while dealing with the members of the family, relatives, friends and the society. Apart from the effect of infertility the couples have an indirect impact on their relation with each other. The low sperm count of the male may not help to conceive, as a result of which the couple is advised ART treatment using donor sperms [4] [5]. ICMR (Indian Council of Medical Research) defines sperm donor as the donor whose sperm sample is collected by the sperm bank with all necessary formalities regulated by ICMR guidelines [1] [6].The sample is processed through several medical tests and procedures and later preserved by the sperm banks. The sperm bank records the donor details viz., their personal, maternal, paternal details along with medical history as per prescribed guidelines. The ART specialist then processes the donated sperm obtained from the sperm bank for the fertility treatment. The selection of the sperm donor profile that matches the male partner of the couple seeking fertility treatment is a complex process. In India, many of the ART specialists specify characteristics of infertile patient based on selected parameters such as blood group, skin color, hair texture, religion etc. and place their requirements to the sperm banks [2]. The sperm bank matches these requirements with the donor profiles stored in their databases and provide the matched donor samples to the ART specialist [5][7]. The couples seeking fertility treatment using donated sperm expect their offspring to have physical appearance similar to them. Favorably, in order to achieve this, it is expected that the sperm donor profile matches with the physical characteristics (attributes) such as skin tone, hair texture, hair color, height, eye color, race etc. of the male partner[8][9]. There is a growing demand from ART specialists and sperm banks for an expert system that will assist them in selecting the most suitable sperm donor profile that fulfills the requirement of the recipient couple. This requirement is the motivation behind developing a mechanism to match donor profiles and will significantly enhance the process of donor selection. This paper presents matching sperm donor profiles using nominal and binary methods. This approach filters the profiles and arranges them in an order of highest matching to the least, based on the physical characteristics of the male partner of the recipient couple.

2. RELATED WORK
The research conducted in the field of fertility is with an objective to improve the success rate and optimize the process of treatment. Many researchers from social sciences , medical sciences, computer sciences, mathematics and statistics has contributed to research in the field of ART, with a prime objective to enhance the ART procedure and optimize the results of treatment cycle [10][11][13-16]. Lucy Frith et. al. surveyed non-biological parents (mothers and fathers) who have used donor conception. The survey addressed some key issues involved in couples conceiving through donated sperm. One of the issues was to study the importance of matching the donor with the non-biological father. The non-biological father preferred his physical attributes to match the donor, whereas low priority was given to interest and occupation [8]. In a study by Jennifer Burr participants perceive the donor's physical characteristics, but also see their husband's physical characteristics in their children [17]. If the offspring's characteristics matches the non-biological father or the couple as a whole, preserves the features of conventional family life and affirms the non-biological father insecurity about his infertility. In some countries sperm donor sample can be selected based on photographs of the donor. But in many countries (including India), law does not permit revealing donor identity and hence this technique cannot be applied at sperm banks operating in such countries. The donor profile matching has to be done on the basis of information recorded as per the guidelines of the regulating authority.

Mathematical techniques such as binary and nominal variable can be used to calculate the similarities between two objects. Irshad Ullah followed binary variable method in an apriori algorithm and experimented on a hospital dataset. The results obtained by performing different experiments proved that such techniques are consistent and may be used to categorize patients for the different treatment purposes [16].

3. METHODOLOGY

3.1 Dataset and Attributes:

This work is based on ten attributes (physical characteristics) selected from those prescribed by ICMR guidelines. Table 1 shows selected attributes along with the abbreviations used later in this paper.

Table 1: Selected Physical Characteristics

Attribute	Abbrev.	Attribute	Abbrev.
Blood Group	BG	Face Shape	FS
Skin Tone	ST	Eye Color	EC
Ethnicity	ETH	Hair Color	HC
Hair Texture	HT	Body Tone	BT
Height	HGH	Nose Shape	NS

Many sperm banks all around the world have made their donor profiles online in an anonymous manner. The identity of the donor is not revealed by any means. Websites such as xytex.com cryobank.com etc. have hundreds of records available for desirous couples or ART specialist to match their requirements. Such anonymous profiles are included in the test dataset and the details are given in table 2. Table 2 also illustrates the number of attributes and total datasets recorded from the respective sperm bank web pages.

Table 2: Dataset from the Sperm Banks

Sperm Banks	Websites	Attrib.	Datasets
California Cryobank	cryobank.com	79	500
Xytex Sperm Bank	xytex.com	101	2100
BabyQuest CryoBank	babyquest.in	69	535
Total Records			3135

3.2 Data Encoding

The algorithms in this paper are implemented using open source programming language R (available for free download at www.r-project.org). R provides wide variety of statistical and graphical techniques; it is highly extensible and is popularly used in scientific experiments. The attributes selected in this work are mostly descriptive in nature and contain textual values. For ease of processing data using R the descriptive data is encoded into suitable numerical form.

Table 3: Numerical Encoding of Attributes

Values / Attrib	1	2	3	4	…
EC	black	brown	amber	dusky	…
ETH	North Indian	South Indian	Asian	European	…
BG	A+	A-	B+	B-	…
HT	smooth	silky	shiny	straight	…
FS	round	oval	pear	heart	…
HGH	< 5.0	5.0 – 5.2	5.3-5.5	5.6-5.8	…
NS	short	flat	straight	pointed	…
ST	light	fair	medium	wheatish	…
BT	weak	healthy	low obesity	high obesity	…
HC	black	brown	burgundy	blonde	…

Table 3 illustrates attributes and their corresponding numerical encodings. A database of 3135 encoded donor records with 10 attributes is used for the sperm donor profile selection. Table 4 illustrates the encoded database.

Table 4: Database of Sperm Donor with Encoded Values

Record No.	BG	ST	ETH	HT	HGH	FS	EC	HC	BT	NS
1	3	5	4	3	4	7	2	1	2	4
2	1	4	1	4	4	3	1	1	3	5
:	:	:	:	:	:	:	:	:	:	:
3135	3	4	3	4	2	3	1	3	3	3

3.3 Nominal Variable Method

The sperm donor profile selection is implemented using a nominal variable or also known as categorical variables of clustering analysis. The nominal variable method computes the dissimilarity between two objects based on the ratio of number of matched attributes of *object a* with *object b*. Here m is the number of matches (i.e., the number of variables for which *object a*, and *object b* are in the same state), and *p* is the total number of variables. The states can be denoted by letters, symbols, or a set of integers, such as *1,2,..,w.*

$$f(a,b,i) = \begin{cases} 1 & a_i = b_i \\ 0 & otherwise \end{cases} \quad \text{..... (1)}$$

Function *f* (*a,b,i*) is the function used for finding the matching attributes between *objects a* and *b*. Here, *i* is the index of matched attributes and varies from *i = 1, 2,…,n*. Weights can be assigned to increase the effect of *m* or to assign greater weight to the matches in variables having a larger number of states. The number of matches i.e., *m* for the attributes of the *object a* and object *b* is computed as shown in equation 2.

$$m = \sum_{i=1}^{n} f(a,b,i) \quad \text{..... (2)}$$

The function *dis()* (equation 3) is dissimilarity function used for matching the features between two objects. The dissimilarity between two objects a, and object b with n attributes can be computed based on the ratio of mismatches.

$$dis(a,b) = (p-m)/p \quad \text{..... (3)}$$

Algorithm: NDonorProfileMatch(D, a, b, c)

Input: D: donor database, a: recipient, b: donor, c: constraint set;

Output: R: result set.

Begin

Step 1: Call *D'= DataReduction(D,c)*

Step 2: For each donor in *D'*

Call *m = DonorAttributeMatch(D',a,b)*for each donor

Calculate the dissimilarity for each donor with respect to recipient record

$dis(a,b) = arg((p-m)/p)$

Step 3: Sort the resultant records obtained in step 2 on dissimilarity values and store in R

Step 4: Return Result dataset R

Stop

Algorithm: DonorAttributeMatch(D',a,b)
Input: D':reduced donor database, a: recipient, b: donor
Output: m: match count an integer

Begin
Step 1: For each attribute i
if (object a_i== object b_i)
m = m + 1
end if
Step 2: Return *m*
Stop

Algorithm: DataReduction(D,c)
Input: D: donor database, c: constraint set;
Output: D': reduced database.

Begin
Step 1: For each i in D
if D_i does not match constraint set c
Add D_i to D'
end if
Step 2: Return *D'*
Stop

3.4 Binary Variable Method

The data matrix also called a two mode matrix (object by variable) is used to represent different entities. The data matrix (fig. 1) represents *n* donors (objects) along with *p* attributes (variables). Attributes here are encoded values of blood group, height, weight, skin tone etc. as given in table 4.

$$\begin{bmatrix} x_{11} & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{nf} & \dots & x_{np} \end{bmatrix}$$

Fig. 1: Data matrix (n X p)

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & : & \\ : & : & : & : \\ d(n,1) & d(n,2) & .. & 0 \end{bmatrix}$$

Fig. 2: Dissimilarity matrix (n X n)

The dissimilarity matrix, also called a one mode matrix (object by object) is a collection of proximities for all pairs of *n* objects. In the dissimilarity matrix (fig. 2) each element of the matrix, *d(a,b)* is the difference between *object a* and *b*. if the value *d(a,b)* is nearer to zero it means that objects *a* and *b* are close to each other, whereas the more it is away from zero it means object *a* and *b* are less similar (differ more). To calculate the dissimilarity between the objects the contingency table $\mathbf{C_{t(a,b)}}$ in table 5 is used.

Table 5: Contingency table (C_t)

		object b		
		1	**0**	**sum**
object a	**1**	*q*	*r*	*q+r*
	0	*s*	*t*	*s+t*
	sum	*q+s*	*r+t*	*p*

Here, *q* is sum of variables that equal 1 for both objects *a* and *b*, *s* is sum of variables that equal 0 for *object a* but equal 1 for *object b*, *r* is sum of variables that equal 1 for *object a* but are 0 for *object b*, and *t* is sum of variables that equal 0 for both *objects a* and *b*. The total of all these four variables is *p*, i.e., $p = q + r + s + t$ [19]. A given binary variable is called as symmetric if it's both states are equally important and carry the similar weight, in other words, it does not matter which outcome should be coded as 0 or 1. Symmetric binary dissimilarity is based on symmetric binary variables. In such case the dissimilarity measure is computed using equation (4), to evaluate the dissimilarity between objects *a* and *b*.

$$dis(a,b) = \frac{r+s}{q+r+s+t} \qquad \dots\dots\dots\dots\dots\dots (4)$$

The distance between two binary variables can also be measured based on similarity instead of dissimilarity using equation (5). The asymmetric binary similarity between the objects *a* and *b*, is calculated as,

$$sim(a,b) = \frac{q}{q+r+s} = 1 - d(a,b) \qquad \dots\dots\dots\dots\dots\dots (5)$$

If the outcomes of the states are not equally valuable then the binary variable is called as asymmetric. Asymmetric binary dissimilarity is based on asymmetric binary variables. Here t, i.e., number of negative matches is ignored (as it is not important) and is calculated using equation (6) [18] [19]

$$dis(a,b) = \frac{r+s}{q+r+s} \qquad \dots\dots\dots\dots\dots\dots (6)$$

The dissimilarity $\boldsymbol{C_t}$ between *object a* (recipient) and *object b* (donor record from D) is computed by measuring the count of variables *q, r, s* as shown in table 6.

Table 6: Dissimilarity (C_t) Calculation

R.No.\ Attributes	**BG**	**ST**	**ETH**	**HT**	**HGH**	**FS**	**EC**	**HC**	**BT**	**NS**
(object a) (recipient record)	3	5	4	3	1	6	2	4	4	5
(object b) i^{th} record from D	4	6	1	3	3	3	2	3	2	4
$\boldsymbol{C_t}$	***s***	***s***	***r***	***q***	***s***	***r***	***q***	***r***	***r***	***r***

Substituting the values of respective variables as computed in table 6 in equation (6):

$$dis(a,b) = \frac{5+3}{2+5+3} = \frac{8}{10} \times 100 = 80\%$$

Such computations will be applied to each record of the dataset and top n records will be selected with the least dissimilarity using the above equation (6). Lesser the computed dissimilarity higher the similarity between the two objects.

Algorithm: BDonorProfileMatch(D, a, b, c)
Input: D: donor database, a: recipient, b: donor, c: constraint set;
Output: R: result set.

Begin
Step 1: Call *D'= DataReduction(D,c)*
Step 2: For each donor in *D'*
Call Ct=*CalculateContingency(D',a,b)* for each donor
Calculate the dissimilarity for each donor with respect to recipient record

$$dj(a,D) = \frac{\sum_{i=1ct:ct=r}^{p} 1 + \sum_{i=1ct:ct=s}^{p} 1}{\sum_{i=1ct:ct=q}^{p} 1 + \sum_{i=1ct:ct=r}^{p} 1 + \sum_{i=1ct:ct=s}^{p} 1}$$

Step 3: Sort the resultant records obtained in step 2 on dissimilarity values and store in R
Step 4: Return Result dataset R

Stop

Algorithm: CalculateContingency(D',a,b)

Input: D':reduced donor database, a: recipient, b: donor

Output: Ct: vector containing contingency values

Begin

Step 1: For each attribute i

if((object a_i == object b_i)&&(a_i==0)&&(b_i!=0))
Ct_i=q
if(object a_i > object b_i)
Ct_i=r
if(object a_i < object b_i)
Ct_i=s
if((object a_i==0) && (object b_i==0))
Ct_i=t

Step 2: Return *Ct*

Stop

---

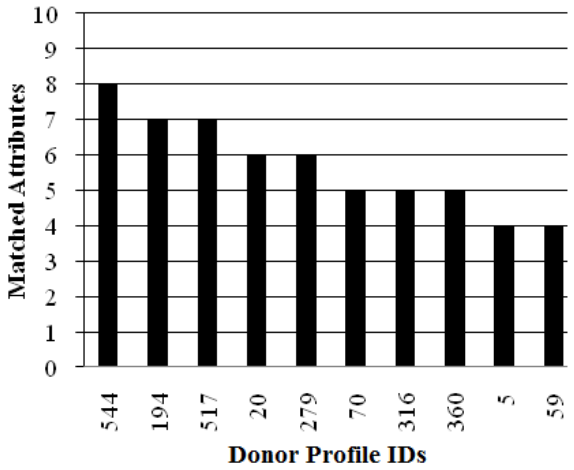
4. RESULTS

Two algorithms first is ***NDonorProfileMatch*** algorithm based on the nominal variable method and second ***BDonorProfileMatch*** on binary variable method respectively are designed and implemented using R. Since, nominal variable is a generalization of the binary variable [19] the results obtained by using both the algorithms are identical. Table 7 depicts the top 10 results for a sample input record (recipient) ***(3,2,1,1,5,4,2,1,2,6)*** used for matching with the donor database.

Table 7: Results using nominal and binary variable method

Attrib's	BG	ST	ETH	HT	HGH	FS	EC	HC	BT	NS	Total Match	Dissimilarity (%) Nominal and Binary Algorithm
Recipient Data / R.No	**3**	**2**	**1**	**1**	**5**	**4**	**2**	**1**	**2**	**6**		
544	**3**	**2**	**1**	**1**	4	**4**	**2**	**1**	**2**	3	8	20.00
194	**3**	**2**	**1**	4	3	**4**	**2**	2	**2**	**6**	7	30.00
517	**3**	**2**	**1**	8	2	**4**	7	**1**	**2**	**6**	7	30.00
20	1	**2**	**1**	**1**	3	**4**	6	**1**	4	**6**	6	40.00
279	**3**	**2**	6	**1**	2	7	**2**	**1**	5	**6**	6	40.00
70	**3**	6	**1**	**1**	3	**4**	7	**1**	1	3	5	50.00
316	**3**	**2**	**1**	6	2	2	7	2	**2**	**6**	5	50.00
360	**3**	**2**	2	**1**	2	**4**	10	**1**	4	5	5	50.00
5	2	4	**1**	**1**	**5**	7	9	2	**2**	4	4	60.00
59	4	1	**1**	6	**5**	**4**	11	2	**2**	1	4	60.00

The results obtained using both the algorithms are based on dissimilarity function i.e. lesser the value of dissimilarity closer is the value to the recipient. Graph 1 depicts the total matched attributes versus the donor profile IDs.



Graph 1: Donor IDs versus Total Attributed Matched

From table 7 and graph 1 it is observed that donor profile 544 matches the recipient the most as compared to the other records. A total of 8 attributes match making it 20% dissimilar. Profiles 194 and 517 both match 7 attributes are the next two matches after 544. In case of profile 544 the height and nose shape does not match, for profile 194 hair color, hair texture and height does not match whereas for profile 517 eye color, hair texture, height does not match with the recipient. However since important features like skin tone, ethnicity, eye color, face shape, body tone etc. have matched record 544 is considered as most suitable profile for the recipient couple.

5. CONCLUSION

The results are satisfactory for donor profile matching based on the selected ten attributes. This technique can be integrated into designing of an expert system/tool, which shall significantly improve the process of selecting the sperm donor profile. The expert system/tool designed using the proposed techniques will assist the fertility specialists, sperm banks and ART patients for finding the best-matching sperm donor profile. However there is large scope for improvement and the results obtained shall act as a motivation. This work is restricted to selected ten attributes of the recipient couple's male partner (non biological father) only. This work has given priority to donor profiles with most matched attributes. However, some profiles with relatively less matches may also be more suitable as the unmatched attributes may exactly match others in the family of the non biological father as compared to the profile with most matches whose unmatched attributes may not match any one from the family of the non-biological father. Further work will be continued with donor profile matching considering additional attributes from sperm donor records including recipients and donors paternal as well as maternal attributes. Attributes other than physical characteristics such as medical history, education etc. shall also be taken into account.

6. ACKNOWLEDGMENTS

The authors sincerely thank the experts Mr. Dilip P. Patil, Director, BabyQuest CryoBank Pvt. Ltd, Mumbai and Dr. Sairaj Bairagi, ART Specialist, Uma Fertility Center, Mumbai who have helped in carrying out this work and providing the necessary support along with testing the outcomes of this work.

7. REFERENCES

[1] Guidelines for ART Clinics in India ICMR/NAMS, 2011, "Introduction, Brief History of ART and Requirements of ART Clinics", pp.1-35.

[2] Fox Dov, 2009, "Racial Classification in Assisted Reproduction", Student Prize Papers, The Yale Law Journal, No. 8, Vol.118 pp. 1844 -1852.

[3] Aghabeigi Narges, Alizadeh S., Saremi A., 2014, "Implementation of Descriptive Algorithm on Infertility Data (Data Mining Case Study)", HealthMed – Vol.8, No .9, pp. 1090 – 1097.

[4] Arthur L. Greil, Kathleen Slauson – Blevins, Julia McQuilan, 2010, "The Experience of Infertility : A Review of Recent Literature", Sociology of Health and Illness, Vol. 32,No. 1, pp.140-62.

[5] Abma JC, Chandra A, Mosher WD., 2005, "Fertility, Family Planning, and Women's Health: New Data From the 2005 National Survey of Family Growth", Vital Health Stat, 23, pp.6 -7.

[6] Susan Koruthu., 2014, "Life Skills Integrated Therapy for Infertility Couple: Case Study", Indian Journal of Applied Research, Volume - 4, Issue - 12, pp. 11 -13.

[7] R.B. Patil and Ajay S. Patil, 2011, "Effective Evaluation and Construction of Data Mining Techniques for Increasing the Success Rate of Procedures Held at ART Clinics", in the proceedings of International Conference on Recent Trends in Information Technology and Computer Science (ICRTITCS-2011), page no.7.

[8] Frith L, Neroli Sawyer, Wendy Kramer, 2012, "Forming a Family with Sperm Donation: A Survey of 244 Non Biological Parents", Reproductive Health Care Ltd., Elsevier, Vol.24, No.7, pp. 709 - 718

[9] Guido Pennings., 2000, "The Right to Choose Your Donor: A Step Towards Commercialization or a Step Towards Empowering the Patient?", Human Reproduction, Vol.15, No.3, pp.508 – 514.

[10] N.M. Chaudhari and B.V. Pawar, 2015, "Microscope Image Processing: An Overview", International Journal of Computer Applications, Vol. 124, No. 12, pp. 23-28

[11] N.M. Chaudhari and B.V. Pawar, 2014, "New Hybrid Approach for Identification of Spermatozoa in Human Semen Sample using Microscope Image Processing Techniques", Advances in Image and Video Processing, Vol. 2, No. 6, pp. 15-24

[12] N.M. Chaudhari and B.V. Pawar, 2013, "Light Scattering Study on Semen Analysis Methods/Techniques", in the Proceedings of Nirma University International Conference on Engineering (NUiCONE), pp. 1-4

[13] M.A. Salam, Joseph Davis, Peter Illingworth and Illingworth, 2005, "Applications of Data Mining Techniques in Assisted Reproductive Technology, in the Proceedings of ACIS, Paper 16.

[14] Jeff Wang and Mark V Sauer, 2006, "In Vitro Fertilization (IVF): A Review of 3 Decades of Clinical Innovation and Technological Advancement, Therapeutics and Clinical Risk Management, Dove Medical Press, Vol : 2, No.4, pp. 355–364.

[15] Paweł Malinowski, Robert Milewski, Piotr Ziniewicz ,Anna Justyna Milewska, Jan Czerniecki, Sławomir Wołczynski, 2014, "The Use of Data Mining Methods to Predict the Result of Infertility Treatment Using the IVF-ET Method", Studies in Logic, Grammar and Rhetoric, Vol. 39, No. 1, pp. 67 – 74.

[16] Irshad Ullah, 2010, "Data Mining Algorithms and Medical Sciences", IJCSIT, Vol 2, No. 6, pp. 127 – 134.

[17] Jennifer Burr, 2009, "Fear Fascination and the Sperm Donor as 'Abjection' in Interviews with Heterosexual Recipients of Donor Insemination", Sociology of Health and Illness, Vol. 16, No.3, pp. 231-245

[18] Nong Ye, 2014, Data Mining Theories, Algorithms and Examples, CRC press, pp. 3-8.

[19] Jiawei Han and Micheline Kamber., 2006, Data Mining Concepts and Techniques, Second Edition, Elsevier, pp. 386-398