Towards a Framework for Privacy Preserving Medical Data Mining Based on Standard Medical Classifications

Aurélien Faravelon[1,2] and Christine Verdier[1]

[1] Laboratoire d'Informatique de Grenoble
Bâtiment IMAG C 220 rue de la chimie, 38400 Saint Martin d'Hères
[2] Groupe de recherche Philosophie, Langage & Cognition
Bâtiment ARSH2 Domaine universitaire 38040 Grenoble Cedex 9
{Aurelien.Faravelon,Christine.Verdier}@imag.fr

Abstract. Privacy-preserving data mining often focuses on data alteration but may bias data patterns interpretation and does not offer different levels of access to patterns according to their use. This paper addresses data mining as a prediction tool and proposes to offer several levels of access to data patterns according to users' trustworthiness. The grounding intuition is that patterns' predictive value depends on their precision that should thus vary according to their use. The following problem is considered: a medical data holder wants to disclose data or data patterns and still control the meaning of the disclosed patterns or of the patterns that may be mined out of the released dataset. To tackle this issue, we propose a framework compliant with existing data mining techniques by modeling trust in terms of data precision and generalising data according to standard medical classifications.

Keywords: Privacy, Data Pattern Hidding, Data Generalisation.

1 Introduction

Privacy, defined as control over one's information, is valued as a means to prevent discriminations for instance and may be protected by hindering the disclosure of individual data. Such an hindrance is particularly needed when it comes to medical data as medical condition influences daily, professional and social life. However, this protection may turn out to be insufficient in the light of predictive techniques such as data mining: not only should data be protected but the predictions mined out of these data, i.e. data patterns should be filtered too to forbid their mischievous use as the possession of data patterns and a few medical data about someone may allow to predict part of their medical future and treat them accordingly.

Tactics such as as data randomisation or the modification of mining processes address this problem but the former bias data patterns interpretation and the later is limited to specific techniques. As data patterns' use strongly depends on their users' identities and may be more or less acceptable, we propose in this

M. Szomszor and P. Kostkova (Eds.): E-Health 2010, LNICST 69, pp. 204–211, 2011.

paper to address data mining's potential threat to privacy by managing the degree of pattern's precision a user is granted access to. Our intuition is that the preciser a pattern is, the more useful, therefore sensitive, it is. Hence, according to their trustworthiness, users may be allowed to access very detailed patterns or only more or less general trends. As a first step in our work, we propose a framework based on data generalisation composed of a trust model, a generalisation algorithm and a set of metrics to assess data distortion. This framework is compliant with existing data mining algorithms and allows to adequately interpret data pattern as the generalisation process, driven by standard medical classifications, safeguards data, and thus patterns, semantics.

The rest of the paper is organised as follow: we first present the works related to our proposition. We then state our problem and outline an algorithm to map one's trustworthiness to data's precision. We eventually present the implementation of our trust model.

2 Related Works

[4] defines knowledge as "valid, novel, potentially useful and ultimately understandable patterns in data" that can "lead to some benefit to the user/task". [9] points out that data mining is valuable in the medical field as elicited data patterns are predictive. [3], for instance, shows that association rules ([1]) mined out of a sample of a nephrology relational database provide insightful diagnosis tools. [3] also highlights the necessary pre-processing of data and the need to mine a large amount of data to get meaningful patterns, thus calling for the development of medical grids to share data and computation power.

The Health-e-child project ([5]), provides such a grid by interconnecting north European pediatric databases. As the grid gathers a wide range of partners with various interests (for example, some are industrials, others public institutions) and goals, it is necessary to filter patterns to prevent their misuse. Two main types of strategies can be identified: authors focus either on the "inference problem", i.e. preventing inference of individual piece of data, or on modifying data to alter patterns's generation. [2] provides an algorithms to hide "sensitive association rules" based on an association rule's support (its frequency) and confidence (the strength of the implication). As [2] mostly defines significance and sensitiveness in terms of support, data are altered to decrease rules' support until a minimal threshold. However, this strategy biases data semantics and the interpretation of data patterns. Furthermore, the "sensitive" rules may not be the most frequent ones that may merely describe the natural course of things. Eventually, such a strategy depends on a specific data mining technique.

K-anonimity, ([11]) addresses the inference problem too as the authors note that even if identifiers are removed from a dataset, a person can be identified if these data are linked to others. K-anonimity guaranties that a person's information will not be distinguishable from at least k-1 other records, thus preventing to identify easily this person. K-anonimity can be achieved by data swapping or attribute generalisation i.e. the downgrading of a value's precision, for instance.

[10] and [12] present algorithms to generalise data. Attributes generalisation, contrary to data swapping or randomisation, safeguards data semantics and thus enable to adequately interpret data patterns. Moreover, k-anonimity may be adapted to any data mining technique: a k-anonymous view on data may be used as a regular input to data mining schemes. However, attributes generalisation neither guaranties that all generalised attributes are at the same level of semantic generalisation nor that at least a certain level of generalisation will be reached and this may interfere with patterns generation. [10] and [12] provide metrics to assess the distorsion resulting from generalisation.

[13] addresses the predictive value of data mining. The authors propose to prevent the generation of association rules whose consequent belongs to the set of items the data holder does not want to be predicted. However, this strategy is only applicable to association rules mining and posits that we already know what must not be predicted.

Eventually, no criterion to chose when patterns should be protected is provided. As data patterns disclosure is a cooperation problem, we propose to use trust as a criterion: the more trustworthy a user is, the more detailed the data patterns they are granted access to may be. Trust is defined in [6] as an expectation of a certain behaviour by someone according to what we know of them and what they do. [7] proposes to compute trust by using a fuzzy cognitive map, a fuzzy graph whose nodes are concepts and edges are causal links. In [6], concepts influencing trust are organised as a graph and the computation propagates causality until a certain level of trust is attained. Thanks to its fuzzyness, the computation can rely on linguistic labels and compute fuzzy causality.

3 A Strategy to Generalise Data and Preserve Privacy

3.1 Problem Statement

Consider the grid provided by the Health-e-Child project: each source holds data and at the level of the grid, several users may mine data. As these users pursue different goals that may be more or less acceptable, we want to offer different levels of access to data patterns according to users' profile, goals and queries. Each user is granted a level of trust, defined as the likelihood to mine data for an acceptable goal. The criteria of this acceptability is defined at the level of the grid, by a consensus among partners, for instance. Therefore, each data holder is able to disclose data so that the retrieved patterns may not be used malevolently, i.e. to map the disclosed data or data patterns to user's trustworthiness by managing patterns' precision. Our solution must be scalable, because of the large amount of data involved, adaptable to various data mining techniques to be easily used and preserve data semantics to enable patterns' interpretation.

To be able to combine our approach with existing distributed mining process, we will henceforth focus on data disclosure at the level of each local source, as these disclosed data may be used as an input for a mining scheme. Table 1 presents a sample of local data from a relational database: each local source possesses such a set of records of the form $<v_1,v_2 \ldots, v_n>$where v_n represents a

value from the domain of definition of an attribute. An association rule such as IF sex=M AND POSTCODE = 02138 THEN disease=A02, with a frequence of 30% can be extracted from this sample.

Values hierarchies can be defined over each attribute's domain of definition. A hierarchy of values is a tree where the root is an empty value and the nodes are the possible values of the domain. Each node represents a preciser value than its parent node, so that the leaves are the precisest values of the domain. Examples of hierarchies are presented on Figure 1 for the domains of definition of the attributes PostCode, Sex and the International Classification of Diseases (ICD). For instance, the hierarchy "PostCodes" presents a tree where the leaves are actual American post codes and their parent nodes are groups of post codes where more and more digits are removed from post code strings. At the top of the tree is the root, an empty value.

Table 1. Patient Records Sample

Sex	DateOfBirth	PlaceOfBirth	Date	DiseaseCode
M	01.02.1990	02138	03.04.2005	A02
M	07.12.1970	02142	03.04.2005	C90
F	23.08.1989	02141	03.04.2005	A11
M	12.06.1954	02138	03.04.2005	A02
F	05.04.1985	02139	03.04.2005	A19
F	04.09.1987	02138	03.04.2005	A02
M	08.11.1993	02142	03.04.2005	B99
M	15.05.1964	02139	03.04.2005	A02
F	30.01.1969	02140	03.04.2005	C90
F	22.03.1957	02138	03.04.2005	C90

Hierarchies allow to define a generalisation function over each domain of definition: a piece of data is generalised when it is mapped to a parent, thus less precise, node.

Definition 1 (Generalisation). *The generalisation, noted* $\{c\} \rightarrow p$ *maps each child value in* $\{p\}$ *to a parent node* p.

As generalisation alters information, we provide a metric to assess data distorsion expressing that the closer we are to the root, the more data are distorted by generalisation. This metrics is inspired by the ones provided in the work on k-anonimity.

Definition 2 (DataDistorsion). *The distorsion of a generalisation mapping a value* v *at level* p *to a value* v' *at level* q *in a hierachy* H *of height* h *with levels 1,2 ... (1 being the less precise level), knowing the weight* $w_{j,j-1}$ *between two levels* j *and* $j-1$ *(with* $2 \leq j \leq h$*) and* $w_{j,j-1}$ *being computed by weight* $(w_{j,j-1})=1/(j_{-1})$*, is defined as: distorsion*$(v,v') = \frac{\sum_{j=q+1}^{p} w_{j,j-1}}{\sum_{j=2}^{h} w_{j,j-1}}$.

Contrary to data randomisation, for instance, data generalisation preserves data semantics as a value does not turn absurd when it is generalised even though the value may be suppressed when mapped to the root. Data can then be correctly interpreted and used as inputs for data mining algorithm. As data generalisation meets two of our criteria (it neither depends on a specific mining scheme nor alters data semantics), we now try to answer our problem by mapping trustworthiness and values hierarchies and thus adapt data generalisation to our purpose.

3.2 Modeling Trust According to Data Precision

Our intuition is that to be useful, patterns have to be precise: very general ones such as "IF smoking THEN throatCancer", even though having a strong confidence and support, are not useful and are not likely to be novel, i.e. to bring the user a potentially harmful piece of knowledge. Therefore, we propose to model trust according to data precision, following this principle: the preciser data is, the more useful and novel patterns are likely to be and then the more trustworthiness is necessary to access them.

To express trust levels, we defined three linguistic labels: "high", "medium" and "low". Drawing on the work on data generalisation, we propose to annotate values hierachies with these linguistic labels. By definition, the root value is the least precise value and then demands the lowest level of trust to be accessed. Each node must have a trust level equivalent or higher than its parents' and some linguistic labels may be skipped. Hierarchies are annotated by business experts and we have posited that at a given level, all sibling nodes have the same trust level.

Figure 1 presents examples of values hierachies annotated with trust levels: the darker the stroke colour is, the higher the trust level required to access data is. On the hierarchy Person, for instance, only two trust levels may be used. We developed a web-based tool to design such hierarchies and our trust model allows us to redefine generalisation to comply with the requirement of a certain level of data generalisation as follow:

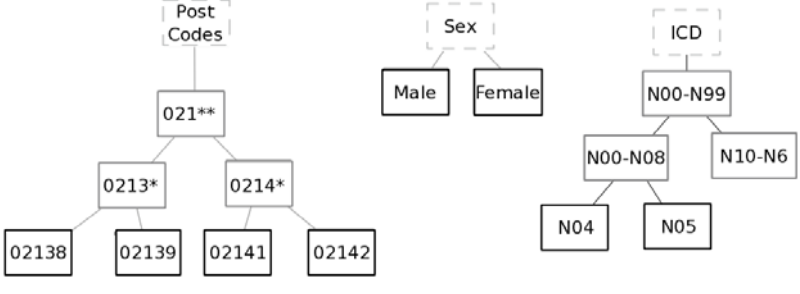

Fig. 1. Hierarchies of Attributes Values

Algorithm 1. Generalisation D' of the dataset D given a hierarchy H and a minimal trust level T

```
for all item d in D do
    d' ← first parent node in H with a necessary trust level at least equal to T
    add d' to D'
end for
return D'
```

Directly applied, this algorithm may lead to over-generalisation if only a certain type of patterns are to be filtered. As a primer approach we propose to generalise only values that can be termed "interesting" according to the considered mining technique. For instance, if only association rules or sequential patterns are to be generalised, as they are mined according to a minimal support, we may only generalise the values satisfying this minimal support. As for classification, feature selection techniques, such as information gain (defined as the number of bits of information obtained for category prediction by knowing the presence or absence of a value), allow to seek values to generalise. Our algorithm may thus be only applied to the subset of values identified as relevant for the mining technique. Table 2 presents Table 1 when generalised (the required level of trust is "medium") for association rules with a support of at least 30%: all the values with a support of at least 30% are to be generalised. From this data set, the pattern IF sex=M AND POSTCODE = 02138 THEN disease=A02 may not be mined any more but a pattern such as IF sex = person AND postcode = "0213*" THEN Intestinal infectious diseases may be mined with a support of 70%. In this case, the support is higher but this pattern is less precise and less usable as the classifiers sex, and postcode are blurred and so is the predicted variable DiseaseCode. We hereafter investigate the implementation of such a generalisation process in the context of data mining.

Table 2. An Example of Generalised Data

Sex	DateOfBirth	PlaceOfBirth	Date	DiseaseCode
Person	01.02.1990	0213*	04.2005	Intestinal infectious diseases
Person	07.12.1970	02142	04.2005	Intestinal infectious diseases
Person	23.08.1989	02141	04.2005	A11
Person	12.06.1954	0213*	04.2005	Intestinal infectious diseases
Person	05.04.1985	02139	04.2005	A19
Person	04.09.1987	0213*	04.2005	Intestinal infectious diseases
Person	08.11.1993	02142	04.2005	B99
Person	15.05.1964	02139	04.2005	Intestinal infectious diseases
Person	30.01.1969	02140	04.2005	Intestinal infectious diseases
Person	22.03.1957	0213*	04.2005	Intestinal infectious diseases

4 Implementation of Our Trust Model

As medical grids have to deal with semantically heterogeneous data because each local source possesses a local data model, we defined a virtual data model (Figure 2), by surveying medical data mining process. The virtual data model therefore regroups data necessary to mine data from a grid of hospitals' databases for instance and represents more or less a patient record, composed of attributes allowing to classify patients and health related events data.

We defined a middleware to map local data to this model: data types are converted and local attributes are mapped to their virtual equivalent. In this model, we expect diseases to be coded according to the ICD. If not, the local classification codes are translated into the ICD codes as most national classifications may be translated to the ICD. The same with the International Nonproprietary Names (INN) for drugs. For each attribute, given that several types of dates (as the date of death, date of birth etc.), for instance, are considered as all being conceptually dates, a values hierarchy is defined if necessary. When standard classifications already exist, we used them as values hierarchies. For instance, the ICD can be considered as such a hierarchy insomuch as it is structured as a tree. The first level of the tree is composed of chapters, that are divided into groups of values (the ICD codes), that may be subdivided in subgroups. Therefore, as we go deeper in the tree, data precision increases and we may thus annotate the ICD with our trust levels as pictured in Figure 1. Our generalisation process is consequently based on a consensual and domain-specific classifications, enabling users to interpret patterns adequately: when a query is issued, a view on data is generated according to the input hierarchies and the user's trust level using the generalisation algorithm we provided. This view may then be fed into a mining scheme. Trust levels are computed by a Fuzzy Cognitive Map, according to rules embedding the definition of goals' acceptability defined by the involved partners and users may be provided with the hierarchies and the data distorsion rate to help them to interpret patterns.

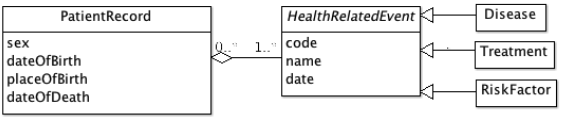

Fig. 2. Virtual Data Model

5 Conclusion

We proposed a strategy driven by medical classifications to provide different levels of access to data patterns. Driven by the intuition that data patterns' value lies in their precision, we proposed to model trust according to data precision and outlined a process to generate views on data according to trust levels. This work

can be regarded as an extension of the work on k-anonimty and both approaches may be merged to provide useful and coherent but filtered data patterns. Further work include investigating features selections techniques to minimise the needed generalisation. We also plan on confronting our trust model to medical experts and are integrating our strategy to a medical grid. This paper was written as part of the STIC-AmSud Project IOSTIC-06 ALAP.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499. Morgan Kaufmann, Santiago de Chile (1994)
2. Dasseni, E., Verykios, V.S., Elmagarmid, A.K., Bertino, E.: Hiding association rules by using confidence and support. In: Moskowitz, I.S. (ed.) IH 2001. LNCS, vol. 2137, p. 369. Springer, Heidelberg (2001)
3. Elfangary, L.M., Atteya, W.A.: Mining databases by means of an incremental association rule learner. In: Proceedings of the 2008 Third International Conference on Convergence and Hybrid Information Technology, pp. 891–896. IEEE Computer Society, Washington, DC, USA (2008)
4. Fayyad, U.M.: Knowledge discovery in databases: An overview. In: Džeroski, S., Lavrač, N. (eds.) ILP 1997. LNCS, vol. 1297, pp. 3–16. Springer, Heidelberg (1997)
5. Freund, J., Comaniciu, D., Ioannis, Y., Liu, P., McClatchey, R., Morley-Fletcher, E., Pennec, X., Pongiglione, G., Zhou, X.: Health-e-child: An integrated biomedical platform for grid-based paediatric applications. CoRR abs/cs/0603036 (2006)
6. Hosmer, L.T.: Trust: The connecting link between organizational theory and philosophical ethics. The Academy of Management Review 20(2), 379–403 (1995)
7. Kim, D.J., Ferrin, D.L., Rao, H.R.: A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. Decis. Support Syst. 44(2), 544–564 (2008)
8. Kosko, B.: Fuzzy cognitive maps. International Journal of Man-Machine Studies 24(1), 65–75 (1986)
9. Palaniappan, S., Awang, R.: Intelligent heart disease prediction system using data mining techniques. In: Proceedings of the 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 108–115. IEEE Computer Society, Washington, DC, USA (2008)
10. Sweeney, L., Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems 10 (2002)
11. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 557–570 (2002)
12. Wang, K.: Bottom-up generalization: a data mining solution to privacy protection. In. In: ICDM, pp. 249–256 (2004)
13. Wang, S.L., Lai, T.Z., Hong, T.P., Wu, Y.L.: Hiding predictive association rules on horizontally distributed data. In: Chien, B.-C., Hong, T.-P., Chen, S.-M., Ali, M. (eds.) IEA/AIE 2009. LNCS, vol. 5579, pp. 133–141. Springer, Heidelberg (2009)