Confidence-Based Reasoning with Local Temporal Formal Contexts*

Gonzalo A. Aranda-Corral[1], Joaquín Borrego Díaz[2], and Juan Galán Páez[2]

[1] Universidad de Huelva, Department of Information Technology,
Crta. Palos de La Frontera s/n. 21819 Palos de La Frontera, Spain
[2] Universidad de Sevilla, Department of Computer Science and Artificial Intelligence,
Avda. Reina Mercedes s/n. 41012 Sevilla, Spain

Abstract. Formal Concept Analysis (FCA) is a theory whose goal is to discover and to extract Knowledge from qualitative data. It provides tools for reasoning with implication basis (and association rules). In this paper we analyse how to apply FCA reasoning to increase confidence in sports betting, by means of detecting temporal regularities from data. It is applied to build a Knowledge based system for confidence reasoning.

1 Introduction

Context modelling and reasoning represents a major paradigm in Artificial Intelligence (AI). It is a useful approach for pragmatic and realistic reasoning in AI. An interesting issue in some types of context reasoning problems is Knowledge's temporal dimension. Knowledge Bases (KB), or databases, may contain information from temporal stamps, bounds or duration. The correctness of reasoning with them depends on a sound selection for time-dependent data (among other features) that will be used in each context. It represents a problem in data mining, particularly for reasoning with association rules, thinking on them as implications with no exact confidence.

Formal Concept Analysis (FCA) [8] is a mathematical theory for data analysis using formal contexts and concept lattices as key tools. Domains can be formally modelled according to the extent and the intent of each formal concept. In FCA, the basic data structure is a formal context (with a qualitative nature) which represents a set of objects and their properties and it is useful both to detect and to describe regularities and structures of concepts. It also provides a sound formalism for reasoning with such structures, mainly Stem Basis and association rules. Therefore, it is interesting to consider its application for reasoning with temporal qualitative data (see e.g. [14]) in order to discover temporal trends.

In this paper, FCA application scope is the challenge of sports betting, specifically, the forecasting of soccer league's results. Forecasting sport results is a fast

* Partially supported by TIN2009-09492 project (Spanish Ministry of Science and Innovation), cofinanced with FEDER founds and *Proyecto de excelencia* TIC-6064 of *Junta de Andalucía.*

J. Cabestany, I. Rojas, and G. Joya (Eds.): IWANN 2011, Part II, LNCS 6692, pp. 461–468, 2011.

growing research area, because of its economic impact in betting markets as well as for its potential application to problems with similar behaviour (markets) [1].

Roughly speaking, three dimensions have been considered for analysing/synthesizing prediction systems: 1)Those which analyse information on teams (endogenous) versus those which analyse results (exogenous); 2)Those which exploit quantitative data versus those which exploit qualitative knowledge, and finally, 3)Statistic-based ones versus other methods. Usually, one can work with hybrid models, and rarely with pure qualitative and exogenous reasoning systems appear in literature, although their use is considered for experiments (for example, frugal methods [3] and based on the recognition heuristic [10]) or as part of hybrid systems (see e.g. [13]). There are two reasons that may justify this point.

On the one hand, transformation from a large quantitative dataset to a qualitative problem is faced with the selection of an acceptable threshold and the discovery of better relations (see e.g. [12]). On the other hand, a qualitative dataset must be accomplished with some amount of information based on confidence, trust or probability of these data sets.

The aim of this paper is to present a method for FCA reasoning on contexts with temporal dimensions that allows the detection of some kind of regularity in data focusing on results from the Spanish soccer league as the source of temporal qualitative information. The method is bet-oriented and its performance is evaluated within a confidence-based reasoning system that increases the number of hits in soccer matches forecasting by using the discovery of temporal trends on data mining and association rules reasoning.

The structure of the paper is as follows. The next section reviews the main features of FCA and association rules on formal contexts. Temporal formal contexts are defined in Sect. 3. The confidence-based reasoning system is described in Sect. 4, and some comments on experimentation are discussed in Sect. 5. Section 6 is devoted to describe future work.

2 Formal Concept Analysis

According R. Wille, FCA mathematizes the philosophical understanding of a concept as a unit of thoughts composed of two parts: the extent and the intent [8]. The extent covers all objects belonging to this concept, while the intent comprises of all common attributes valid for all the objects under consideration. It also allows the computation of concept hierarchies from data tables. In this section, we succinctly present basic FCA elements, although it is assumed that the reader is familiar with this theory (the fundamental reference is [8]).

We represent a formal context as $M = (O, A, I)$, which consists of two sets, O (objects) and A (attributes) and a relation $I \subseteq O \times A$. Finite contexts can be represented by a 1-0-table (representing I as a Boolean function on $O \times A$). The FCA main goal is the computation of the concept lattice associated to the context. In this paper it works with logical relations on attributes which are valid in the context. For $X \subseteq O$ and $Y \subseteq A$ we can define

$$X' := \{a \in A \mid oIa \text{ for all } o \in X\} \quad Y' := \{o \in O \mid oIa \text{ for all } a \in Y\}$$

Logical expressions in FCA are *implications between attributes*, pair of sets of attributes, written as $Y_1 \to Y_2$, which is true with respect to $M = (O, A, I)$ according to the following definition. A subset $T \subseteq A$ *respects* $Y_1 \to Y_2$ if $Y_1 \not\subseteq T$ or $Y_2 \subseteq T$. It says that $Y_1 \to Y_2$ holds in M ($M \models Y_1 \to Y_2$) if for all $o \in O$, the set $\{o\}'$ respects $Y_1 \to Y_2$. In that case, $Y_1 \to Y_2$ is *an implication* of M.

Definition 1. *Let $\mathcal{L}$ be a set of implications and L an implication of M.*

1. *L follows from $\mathcal{L}$ ($\mathcal{L} \models L$) if each subset of A respecting $\mathcal{L}$ also respects L.*
2. *$\mathcal{L}$ is complete if every implication of the context follows from $\mathcal{L}$.*
3. *$\mathcal{L}$ is non-redundant if for each $L \in \mathcal{L}$, $\mathcal{L} \setminus \{L\} \not\models L$.*
4. *If $\mathcal{L}$ is a basis for M is complete and non-redundant.*

It can obtain a basis from the *pseudo-intents* [11] called *Stem basis*:
$\mathcal{L} = \{Y \to Y'' \ : \ Y \text{ is a pseudointent}\}$

The so-called *Armstrong rules* provides an implicational reasoning:

$$R1 : \frac{}{X \to X} \quad R2 : \frac{X \to Y}{X \cup Z \to Y} \quad R3 : \frac{X \to Y, \ Y \cup Z \to W}{X \cup Z \to W}$$

Let $\vdash_A$ be the proof relation by Armstrong rules. It holds that implicational bases are $\vdash_A$-complete [6]: If $\mathcal{L}$ is a implicational basis for M, and L an implication, then $M \models L$ if and only if $\mathcal{L} \vdash_A L$.

In order to work with formal contexts, stem basis and association rules, the Conexp[1] software has been selected. It has been used as a library to build the module which provides implications (and association rules) to the reasoning module. This module is a production system based on which was designed for [4]. It works with Stem Basis, and entailment is based on the following result.

Theorem 1. *Let $\mathcal{S}$ be a stem basis associated with the context M, o a new document tagged with $A_1, \ldots, A_n$. The following conditions are equivalent:*

1. *$\mathcal{S} \cup \{A_1, \ldots A_n\} \vdash_p Y$ ($\vdash_p$ is the entailment from the production system).*
2. *$S \vdash_A A_1, \ldots A_n \to Y$*
3. *$M \models \{A_1, \ldots A_n\} \to Y$.*

We can consider a Stem Basis as an adequate production system in order to reason and predict results. However, Stem Basis is designed for entailing true implications only, without any exceptions into the object set nor implications with a low number of counterexamples in the context. Another more important question is about predictions, we are interested in obtaining some methods for selecting a result among all obtained results (even if they are mutually incoherent), and theorem 1 does not provide such a method. Therefore, it is better to consider rules with confidence instead of true implications and the initial production system must be revised for working with confidence.

Researching on sound logical reasoning methods with association rules is a relatively recent research line with promising applications [7]. In FCA, association rules are implications among sets of attributes. Confidence and support are

[1] `http://sourceforge.net/projects/conexp/`

defined as usual. Recall that the *support* of X, $supp(X)$, of a set of attributes X is defined as the proportion of objects which satisfy every attribute of X, and the *confidence* of an association rule is $conf(X \Rightarrow Y) = supp(X \cup Y)/supp(X)$. Confidence can be interpreted as an estimate of the probability $P(Y|X)$, the probability of an object satisfying every attribute of Y under the condition that it also satisfies every one of X. Conexp software provides the association rules, as well as, their confidence for contexts.

3 Data and Temporal Contexts

A temporal context on a set of objects is defined as follows:

Definition 2. *Let O be a set of objects.*

1. *A* **temporal context** *on O is a context $M = (O_1, A, I)$ where $O_1 \subseteq O \times \mathbb{N}$*
2. *A* **contextual selection** *is a map $s : O \to \mathcal{P}(O_1) \times \mathcal{P}(A)$*
3. *A* **contextual KB for an object** *o w.r.t. a selection s with confidence γ is a subset of association rules with confidence greater or equal that γ of the formal context associated to $s(o) = (s_1(o), s_2(o))$, that is, to the context*

$$M(s(o)) := (s_1(o), s_2(o), I_{\upharpoonright s_1(o) \times s_2(o)})$$

In this paper only set of association rules extracted by Conexp with confidence greater than a threshold γ for a contextual selection are used as contextual KB.

3.1 Temporal Contexts for Soccer League

For both selecting data and building contexts, some assumptions on forecasting in soccer league matches have been considered. Reconsiderations of such decisions can be easily computed in the system. First, we consider that the regularity of team's behaviour only depends on the contextual selection that has been considered. This contextual selection is obtained by taking matches from the last X weeks backwards, starting from the week just before the one we want to forecast. Second, since FCA methods are used to discover regularity features, thus it does not consider forecasting exceptions (unexpected results). Therefore, the model can be considered as a starting point for betting expert who would adjust attributes, in order to more personalised criteria.

These attributes have to be computed and used to entail the forecasting. This analysis is assisted by Conexp. ConExp software is used to compute and analyze the concept laticces associated to the temporal contexts. In this way, the expert can evaluate the goodness of the attributes (and the thresholds defining them) (See Fig. 1). The attribute $ID\_1\_T\_16$ is defined by: 'the budget of $team_2$ is greater than γ_1 times the budget of $team_1$', where γ_1 is the threshold the expert must estimate. In the concept lattice we can observe that the biggest concept containing the attributes $team_2\_wins$ and $ID\_1\_T\_16$ covers the about the 10% of the objects owned by the first attribute, therefore it is suggested to use the second attribute for reasoning with association rules to get a prediction.

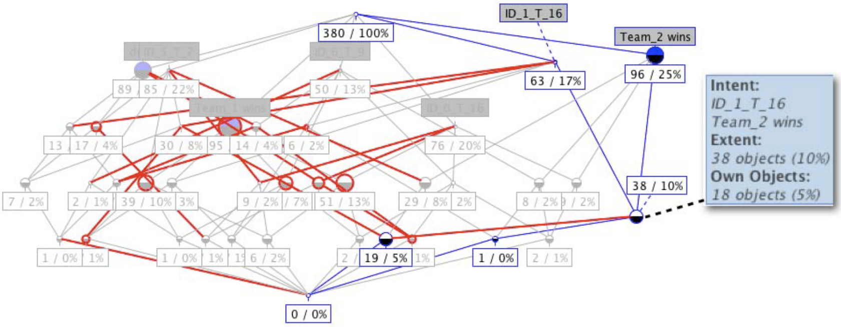


Fig. 1. Concept Lattice for the match *Málaga-Sevilla* (week 31, season 2009-10)

The system computes the value of an amount of attributes on objects. Experimentally a boolean combination of attributes is possible. Once the temporal context has been computed, the system can build contextual selections by selecting the match and the attribute set. The selection of attributes was made by considering four kinds of factors: those related with the classification, the history of teams' matches in the recent past, results of direct matches and other non related results, as for example the difference between team budgets. Seventeen relevant attributes were selected.The attribute set has three special attributes, $Team_1$ wins (1), $Team_2$ wins (2) and draws (X).

4 Confidence-Based Reasoning System

The reasoning system works on facts of the type (a, c), where a is an attribute and c is the estimated probability of the trueness of a, which we also call confidence (by similarity with the same term for association rules).

The system has a module for a confidence-based reasoning system (Fig. 2). Its entries for a match $Team_1$ - $Team_2$ are: the contextual Knowledge basis for a threshold given as rule set and attribute values for the current match

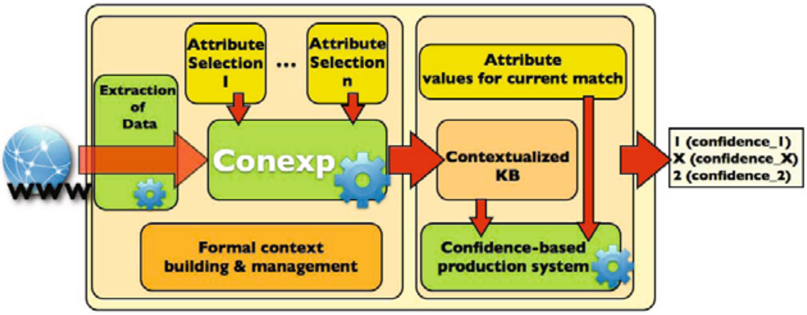
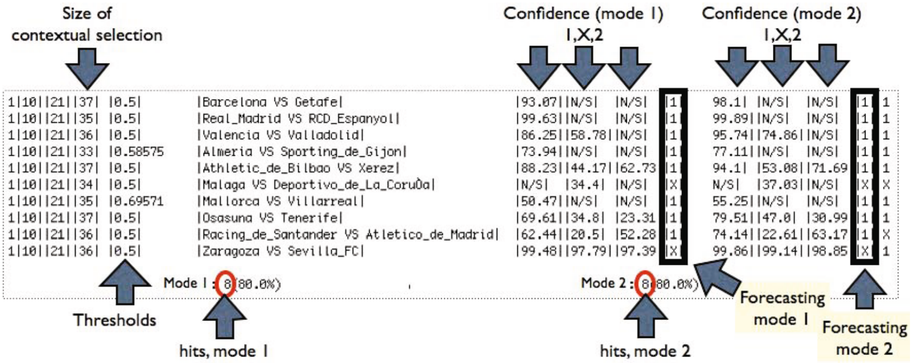


Fig. 2. Context based reasoning system


Size of contextual selection
Confidence (mode 1) 1,X,2
Confidence (mode 2) 1,X,2
Thresholds
hits, mode 1
hits, mode 2
Forecasting mode 1
Forecasting mode 2

```
1|10||21||37| |0.5|      |Barcelona VS Getafe|                    |93.07||N/S|  |N/S|  |1   98.1| |N/S|  |N/S|  |1  1
1|10||21||35| |0.5|      |Real_Madrid VS RCD_Espanyol|            |99.63||N/S|  |N/S|  |1   99.89||N/S|  |N/S|  |1  1
1|10||21||36| |0.5|      |Valencia VS Valladolid|                 |86.25||58.78||N/S|  |1   95.74||74.86||N/S|  |1  1
1|10||21||33| |0.58575   |Almeria VS Sporting_de_Gijon|           |73.94||N/S|  |N/S|  |1   77.11||N/S|  |N/S|  |1  1
1|10||21||37| |0.5|      |Athletic_de_Bilbao VS Xerez|            |88.23||44.17||62.73 |1   94.1| |53.08||71.69 |1  1
1|10||21||34| |0.5|      |Malaga VS Deportivo_de_La_CoruÒa|       |N/S|  |34.4| |N/S|  |X   N/S|  |37.03||N/S|  |X  X
1|10||21||35| |0.69571   |Mallorca VS Villarreal|                 |50.47||N/S|  |N/S|  |1   55.25||N/S|  |N/S|  |1  1
1|10||21||37| |0.5|      |Osasuna VS Tenerife|                    |69.61||34.8| |23.31 |1   79.51||47.0| |30.99 |1  1
1|10||21||36| |0.5|      |Racing_de_Santander VS Atletico_de_Madrid| |62.44||20.5| |52.28 |1   74.14||22.61||63.17 |1  X
1|10||21||36| |0.5|      |Zaragoza VS Sevilla_FC|                 |99.48||97.79||97.39 |X   99.86||99.14||98.85 |X  1

              Mode 1 : 8(80.0%)                              Mode 2 : 8(80.0%)
```

Fig. 3. Forecasting results screenshot

(except 1,X,2) as facts, all of them with a confidence (whose value depends on the reasoning mode, see below). The production system is executed and the output is a triple $<(1, c_1), (X, c_x), (2, c_2)>$ of attribute, confidence for this match. The attribute with greater confidence is selected as the prediction.

The execution of the production system is as regular. There exist several modes of confidence computing of facts, which are based in uncertain reasoning in Expert Systems [9]. Any attribute/fact a is initialized with confidence

$$conf(a) := \frac{|\{o \ : \ oIa\}| + 1}{|O| + 1}$$

The most promising computation modes used are:

Mode 1: As usual in Expert systems: If rule $r : \{a_1, \ldots, a_k\} \to \{c_1, \ldots, c_n\}$ with confidence rule $conf(r)$ is fired on the facts $(a_1, conf(a_1)), \ldots, (a_k, conf(a_k))$, the confidence estimated for each c_i by the rule is

$$conf_n(c_i) = conf(r) \cdot \min(conf(a_1), \ldots, conf(a_k))$$

and it updates $conf(c_i)$ as $conf(c) := conf(c) + conf_n(c) \cdot (1 - conf(c)$.

Mode 2: If c is obtained by firing the rule r, define $conf(c) = P(c) \cdot Q(c)$ where

$$P(c) := f_p(c, r) = conf(r), \ \ Q(c) := f_q(c, r) = \min(conf(a_1), \ldots, conf(a_k))$$

If it entails c by firing of other rule r' produces the update of $conf(c)$ by updating

$$P(c) := P(c) + f_p(c, r') - P(c) \cdot f_p(c, r'), \quad Q(c) := Q(c) + f_q(c, r') - Q(c) \cdot f_q(c, r')$$

With respect to the thresholds for confidence, currently the system allows the user to select them by hand or by using the automatic selection mode which is:

$$\gamma = \max(\{conf(a) \ : \ \text{exists } \{\emptyset\} \to Y \text{ rule of the KB s.t. } a \in Y\} \cup \{0.5\})$$

Fig. 3 shows forecasting for week 21 of the Spanish premier league (2009-10).

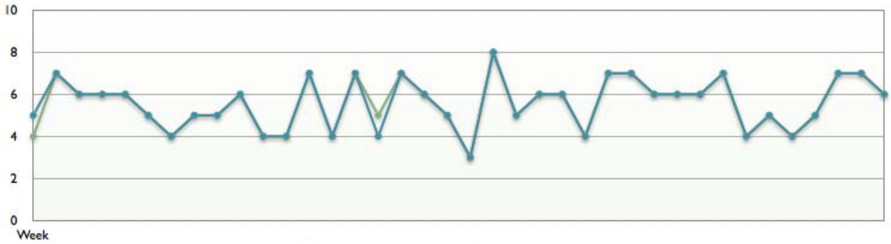


Fig. 4. Hits on 2009-10 league

5 Experiments

It ran an experiment for the Spanish premier and second division soccer leagues from 2009-10. In Fig. 4 hits for premier league are graphically depicted.

About data source, temporal contexts for forecasting results were built by data extracted from the RSSSF Archive (http://www.rsssf.com). Objects are matches (with temporal stamp (week, year)) and attributes are computed. Data was collected for the past four years. The size of the temporal context is about 300 objects and 17 attributes (although several of them are parametrized, i.e. , ranking difference above a threshold). Thus, $|I|$ is about 5,100 pairs.

Experiments with the system show forecasts of about 57.37% in mode 2 and 56.32% in mode 1 by a selection of ten qualitative attributes and contextual selection based on the previous 38 matches of each team (Fig. 4). Such an percentage for a qualitative reasoning system may be considered as an acceptable result comparable with expectable results of experts [3].

It is interesting to comment that the contextual selection for the premier league is not the best for the second league that we have found. For the second league is better to consider complete sets of results. The results, under the conditions of official spanish bet system are: three awards are achieved: 583.42 eur. of earning, with a cost of 38 eur. corresponding to 76 bets (two bets by week).

6 Concluding Remarks and Future Work

A confidence-based reasoning system that works on sub-contexts extracted from a temporal formal context, built for soccer bets, is presented. The system has some similarities with [13], although the reasoning system based on FCA is qualitative while the cited system is hybrid (bayesian reasoning). Pure qualitative reasoning was selected based on the aim of discovering trends (under a contextual selection) represented in the form of association rules with high confidence. It is worth noting that due to the proprietary nature of prediction models, it is difficult to compare them with our system.

Part of our ongoing work includes three research lines. Firstly, it is more interesting to apply methods for the automated definition of new attributes [2]. Secondly, since Attribute logic based on implications does not suffer from inconsistencies (two mutually results can be derived), it was necessary to select the attribute with higher confidence. However, it seems more sound to decide this by using more sophisticated methods. And, finally, the selection of thresholds can be refined to achieve a better dependence among attributes, for this, methods in data mining could be used (see e.g. [12]).

With respect to computational features, computing tasks are feasible (due to the relatively small data size). However, the summation of additional data and attributes could make it necessary to apply conservative retraction methods [5,2] to work with a contextual KB of a feasible size.

References

1. Why Spain will win..., Engineering & Technology (June 5–18, 2010)
2. Alonso-Jiménez, J.A., Aranda-Corral, G.A., Borrego-Díaz, J., Fernández-Lebrón, M.M., Hidalgo-Doblado, M.J.: Extending Attribute Exploration by Means of Boolean Derivatives. In: Proc. 6th Int. Conf. Concept Lattices and Their Applications (CLA 2008), pp. 121–132 (2008)
3. Andersson, P., Ekman, M., Edman, J.: Forecasting the fast and frugal way: A study of performance and information-processing strategies of experts and non-experts when predicting the World Cup 2002 in soccer, Working Paper Series in Business Administration 2003:9, Stockholm School of Economics (2003)
4. Aranda-Corral, G.A., Borrego-Díaz, J.: Reconciling Knowledge in Social Tagging Web Services. In: Corchado, E., Graña Romay, M., Manhaes Savio, A. (eds.) HAIS 2010. LNCS(LNAI), vol. 6077, pp. 383–390. Springer, Heidelberg (2010)
5. Aranda-Corral, G.A., Borrego-Díaz, J., Fernández-Lebrón, M.M.: Conservative Retractions of Propositional Logic Theories by Means of Boolean Derivatives: Theoretical Foundations. In: Carette, J., Dixon, L., Coen, C.S., Watt, S.M. (eds.) MKM 2009, Held as Part of CICM 2009. LNCS, vol. 5625, pp. 45–58. Springer, Heidelberg (2009)
6. Armstrong, W.: Dependency structures of data base relationships. In: Proc. of IFIP Congress, Geneva, pp. 580–583 (1974)
7. Balcázar, J.L.: Redundancy, Deduction Schemes, and Minimum-Size Bases for Association Rules. Logical Methods in Computer Science 6(2), 1–23 (2010)
8. Ganter, B., Wille, R.: Formal Concept Analysis - Mathematical Foundations. Springer, Heidelberg (1999)
9. Giarratano, J.C., Riley, G.D.: Expert Systems: Principles and Programming. Brooks/Cole Publishing Co., Pacific Grove (2005)
10. Goldstein, D.G., Gigerenzer, G.: Models of ecological rationality: the recognition heuristic. Psychological review 109(1), 75–90 (2002)
11. Guigues, J.-L., Duquenne, V.: Familles minimales d' implications informatives resultant d'un tableau de donnees binaires. Math. Sci. Humaines 95, 5–18 (1986)
12. Imberman, S.P., Domanski, B., Orchard, R.A.: Using Booleanized Data To Discover Better Relationships Between Metrics. In: Int. CMG Conference, pp. 530–539 (1999)
13. Min, B., Kim, J., Choe, C., Eom, H., McKay, R.I.: A compound framework for sports results prediction: A football case study. Know.-Based Syst. 21(7), 551–562 (2008)
14. Neouchi, R., Tawfik, A.Y., Frost, R.A.: Towards a Temporal Extension of Formal Concept Analysis. In: Proc.14th Conf. Canadian Soc. on Comp. Studies of Intell. LNCS, vol. 2056, pp. 335–344. Springer, Heidelberg (2001)