Selecting Attributes for Sport Forecasting using Formal Concept Analysis

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

[1]Department of Information Technology, Universidad de Huelva, Spain
gonzalo.aranda@dti.uhu.es
[2]Department of Computer Science and Artificial Intelligence, Universidad de Sevilla, Spain
jborrego@us.es, juangalan@us.es

Abstract

In order to address complex systems, apply pattern recognition on their evolution could play an key role to understand their dynamics. Global patterns are required to detect emergent concepts and trends, some of them with qualitative nature. 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). Implications and association rules are usefull to reasoning on previously selected attributes, providing a formal foundation for logical reasoning. 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.

Introduction

Formal Concept Analysis (FCA) (Ganter & Wille 1999) 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 in order to discover temporal trends (Aranda-Corral et al. 2011).

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 growing research area, because of its economic impact in betting markets as well as for its potential application to problems with similar behaviour (markets) (Inst. Engineering and Technology 2010). Considering sports betting as a complex system, soccer leagues represent a challenging system with a huge amount of knowledge, available through WWW, and its behaviour is weekly exhaustive analysed by journalists, betting companies and supporters. 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 (Goldstein & Gigerenzer 2009) and based on the recognition heuristic (Goldstein & Gigerenzer 2002)) or as part of hybrid systems (see e.g. (Min et al. 2008)). 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. (Imberman et al. 1999)). 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 describe all researching work made for selecting and computing attribute sets related to soccer results, into a specific framework: FCA, and starting from soccer match results, with no previous analysis of any other specific attributes. This task is previous to build an Expert System for advising sport betting which could detects some kind of regularities on data. Concept lattices, which are computed from attribute values, represent a mathematical structure of relationships among the concepts which are involved in selected sport events to study. Since this method is bet-oriented, its performance is evaluated within a confidence-based reasoning system. This sistem increases number of hits in soccer matches forecasting, discovering temporal trends by means of data mining and association rules reasoning. The analysis of attributes has been used in (Aranda-Corral et al. 2011) to describe a confidence-based (and contextual) reasoning system for forecasting sports betting. In this paper we analyse the attribute selection problem as a problem of selection of features that shape the behaviour

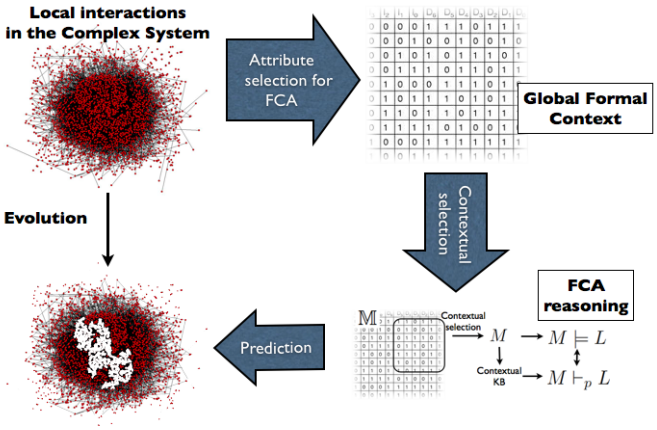


Figure 1: FCA based model for prediction of qualitative features of Complex Systems

of the complex system that represents professional soccer leagues. Theoretical framework, on which this model is based on, will be presented at (Aranda-Corral et al. 2011b). Due to a really huge amount of information, attribute selection advised by experts is mandatory. In fact, the system can be considered as a reasoning model based on bounded rationality and recognizition heuristics. and focused on features which were considered as important by human experts. Therefore, the system aims to forecast results, but it is designed based on bounded rationality models, instead of statistic models (although in the future hybrid models will be considered.

The system is a first prototype from a more general system, which are building to analyse qualitative features of Complex Systems (see Fig. 1), using FCA. The idea is to isolate qualitate attributes from (past) local interactions among components of complex system and to apply FCA tools in order to predict properties system's behavior in a near future.

Background: Formal Concept Analysis

According to R. Wille, FCA (Ganter & Wille 1999) mathematizes the philosophical understanding of a concept as a unit of thoughts composed of two parts: the extent and the intent. 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 (the fundamental reference is (Ganter & Wille 1999)).

A formal context $M = (O, A, I)$ 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 (identifying I with a Boolean function on $O \times A$). See Fig. 2 for an example of formal context about live beings.

The FCA main goal is the computation of the concept lattice associated to the context. Given $X \subseteq O$ and $Y \subseteq A$ it defines

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

A (formal) concept is a pair (X, Y) such that $X' = Y$ and $Y' = X$. For example, concepts from living beings formal context (Fig. 2, left) is depicted in Fig. 2, right.

Using this Fig. 2, each node is a concept, and its intension (or extension) can be formed by the set of attributes (or objects) included along the path to the top (or bottom). E.g. The node tagged with the attribute Legs represents to the concept $(\{Legs, Mobility, NeedWater\}, \{Cat, Frog\})$.

In this paper it works with logical relations on attributes which are valid in the context. Logical expressions in FCA are *implications between attributes*. An implication is a 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, it is said that $Y_1 \to Y_2$ is *an implication* of M.

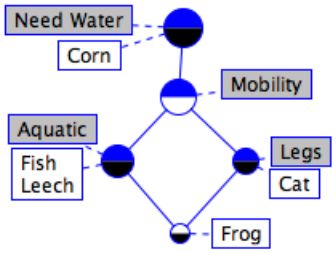
Definition. 1 *Let $\mathcal{L}$ be a set of implications and L be an implication.*

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. *$\mathcal{L}$ is a (implication) basis for M if $\mathcal{L}$ is complete and non-redundant.*

It can obtain a basis from the *pseudo-intents* (Guigues & Duquenne 1986) called *Stem Basis* (SB):

$$\mathcal{L} = \{Y \to Y'' \ : \ Y \text{ is a pseudointent}\}$$

A	B	C	D	E
	Need water	Aquatic	Mobility	Legs
Cat	X		X	X
Leech	X	X	X	
Frog	X	X	X	X
Corn	X			
Fish	X	X	X	



1 < 5 > { } ==> Need water;
2 < 3 > Need water Aquatic ==> Mobility;
3 < 2 > Need water Legs ==> Mobility;

Figure 2: Formal context, associated concept lattice and Stem Basis

A SB for the formal context on live beings is provided in Fig. 2 (right). It is important to remark that SB is only an example of a basis for a formal context. In this paper any specific property of the SB can be used, and it can be replaced by any implication basis.

It is possible to extend $\models$ in relation to any propositional formula with propositional variables in A, by considering each object $o \in \mathbb{M}$ as a valuation v_o on $\mathbb{A}$ defining

$$v_o(A) = 1 \Longleftrightarrow (o, A) \in \mathbb{I}$$

Thus $M \models F$ if and only if for any $o \in O$ it holds that $v_o \models F$.

The *Armstrong rules* (Armstrong 1974) provides a formal basis for implicational reasoning:

$$\frac{}{X \to X} \quad \frac{X \to Y}{X \cup Z \to Y}, \quad \frac{X \to Y, Y \cup Z \to W}{X \cup Z \to W}$$

A set of implications is closed if and only if the set is closed by these rules (Armstrong 1974). By defining $\vdash_A$ as the proof relation by Armstrong rules, it holds that the implicational bases are $\vdash_A$-complete:

Theorem 2 *Let $\mathcal{L}$ be an 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 is used as a library to build the module which provides the implications (and association rules) to the reasoning module of our system. The reasoning module is a production system based on which was designed for (Aranda-Corral & Borrego-Díaz 2010). Initially it works with SB, and entailment is based on the following result:

Theorem 3 *Let $\mathcal{L}$ be a basis for M and $\{A_1, \ldots, A_n\} \cup Y \subseteq A$. The following conditions are equivalent:*

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

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

Association rules for a a formal context

We can consider a Stem Basis as an adequate production system in order to reason. 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 arises when it works on predictions. In this case we are interested in obtaining methods for selecting a result among all obtained results (even if they are mutually incoherent), and theorem 3 does not provide such a method. Therefore, it is better to consider association rules (with confidence) instead of true implications and the initial production system must be revised for working with confidence.

Researching on logical reasoning methods for association rules is a relatively recent promising research line (Balcázar 2010). In FCA, association rules are implications between sets of attributes. Confidence and support are 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 a association rule is $conf(X \to Y) = supp(X \cup Y)/supp(X)$. 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 association rules (and their confidence) for formal contexts.

Reasoning under contextual selection. Logical Foundations

The model (described in (Aranda-Corral et al. 2011b)) is composed of events (objects) which have a number of properties (attributes). They consitute a *universal formal context* $\mathbb{M}$ (which we call *monster context* following the tradition in Model Theory). Thus $\mathbb{M}$ can be considered as the *global memory* from which subcontexts are extracted. Once the specific context is considered, it is also possible to consider background knowledge Δ (in form of propositional logic formulas) which would be combined with the knowledge extracted from formal context (Stem basis or association rules).

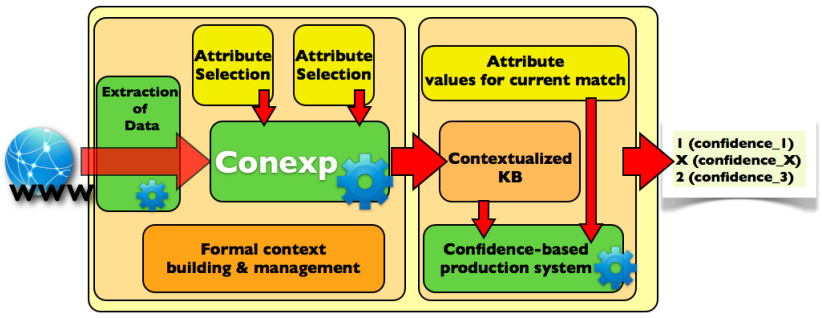


Figure 3: Context based reasoning system

Definition. 4 *Let $\mathbb{M} = (\mathbb{O}, \mathbb{A}, \mathbb{I})$ be the monster context, and let O be a set of objects.*

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

Contextual KBs is useful for entailing attributes on an object. The reasoning model on $\mathbb{M}$ is argumentative, where the argument is based on KBs extracted from subcontexts (Aranda-Corral et al. 2011b):

Definition. 5 *Let L be an implication and Δ a background knowledge. It is said that L is a possible consequence of $\mathbb{M}$ under Δ, $\mathbb{M} \models^{\Delta}_{\exists} L$, if there exists M a nonempty subcontext of $\mathbb{M}$ such that $M \models \Delta \cup \{L\}$.*

Note that by theorem 3, when Δ is a set of implications, it holds that $\models_\exists$ is equivalent to $\vdash_\exists$ which is defined by: $\mathbb{M} \vdash_\exists L$ if there exists $M \models \Delta$ a subcontext of $\mathbb{M}$ such that $S \vdash_p L$ (where S is a stem basis for M).

The role of attribute selection for formal contexts

Attributes are essentials in the contextual selection to build good formal contexts. Association rules are extracted from the contexts and those are used by the production system. By means of these association rules and some initial facts based on the match we want to forecast the production system infers the confidence (probability) for each one of the three possible results of a match, home team wins, draw or away team wins. Thus attributes constitute one of the most important and sensitive parts of the system. They are sensitive because on how they represent the behavior of the teams will depend the accuracy of the inferred results.

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). See (Aranda-Corral et al. 2011) for a more detailed description of the reasoning system.

The system has a module for a confidence-based reasoning system (Fig. 3). 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 (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. Production system execution is standard, with several modes for confidence computing of results based in uncertain reasoning in Expert Systems (Giarratano & Riley 2005). Any attribute/fact a is initialized with confidence

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

Attributes and formal 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.

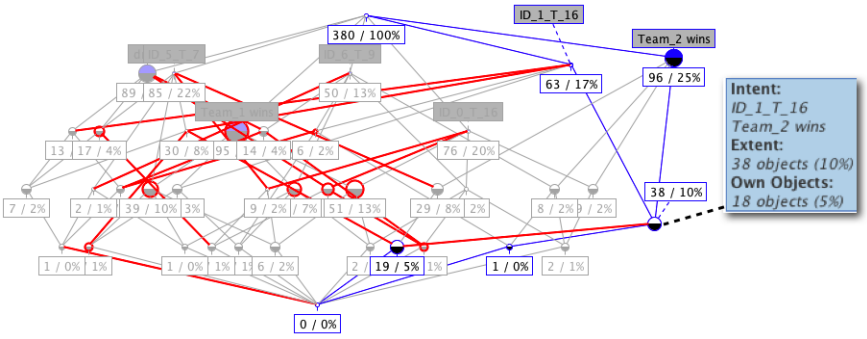


Figure 4: Concept Lattice for the match *Málaga-Sevilla* (week 31, season 2009-10)

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 order to select most interesting attributes for the system, starting from an initial configuration, user can compute the associated concept lattice and check it. In this way, attributes goodness (and thresholds) can be evaluated to reconsider current attribute selection. For example, in Fig. 4, the concept lattice associated to contextual selection for *Málaga-Sevilla* match is shown. This contextual selection is obtained from a given attribute selection and last 38 weeks matches before. In this concept lattice, 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.

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).

With respect to data source, they are automatically extracted from RSSSF Archive[2]. Objects are matches and attributes are a list of features, including temporal stamp (week, year). Data was collected for the past four years. Actually the size of the context is about 300 objects and 18 attributes (although several of them are parametrized, see section bellow). Thus, $|I|$ is about 5,100 pairs.

[2]http://www.rsssf.com

Attribute selection

We have chosen a small set of attributes with many possibilities through a few customizable parameters. When these parameters are having set up with proper values, the set of attributes will represent team's logical behavior.

Recall that formal concept analysis works with qualitative attributes and all teams information which we work with are quantitative data. Thus it is necessary to convert quantitative attributes into qualitative ones. This task is left to users by choosing a proper threshold to each attribute.

Before choosing the set of base attributes , we have carried out a analysis on information about soccer results . The aim have been to discover which factors are more influential in teams behavior and which ones are less influential. First of all, we have collected any interesting factor found, and after analyzing each one, individually, we have chosen most suitable ones. Examined factors can be classified in four different categories (see Table 1): those related to season's classification, those related to previous team's results, those related to historical direct matches and any other factors. It is worth to note that to increase possibilities of the attribute set, and considering the Boolean nature of formal context attributes, we have added the option to create new ones by means of logical combinations of these attributes.

According to considered factors, the system computes a base set of 18 attributes, which are customizable by some parameters. This will let us to obtain a diverse set of attributes. In Table 2 attributes are specified. Four parameters are used:

- Threshold: Parameter to be used to translate quantitative attribute values into qualitative ones.
- Team: Recall that in the formal context considered, objects are matches but attributes belongs to team properties. This parameter will set the team from object (match) on which attribute will be considered. It has two possible values: {HOME, AWAY}. Thus, usually, we will have

twice each attribute at context, once for home team and once for away.

- Number of Matches: sets the number of past matches to be considered when some attributes are computed, e.g. the ones associated to previous team results.

- Kind of matches: sets past matches type to be taken into account to compute some attributes, considering home/away team's condition at matches. Three possible values: {MATCHES AS HOME TEAM, MATCHES AS AWAY TEAM, ALL MATCHES}

With these parameters, and the possibility to compound attributes, it is possible to build a detailed attributes set. Note that experiments show that simplest and most logical attributes give a good team behavior representation. Although we consider that a versatile attributes set, as above described, was necessary because of a huge number of factors can determine the result of a soccer match. Task of customizing the attribute set is left to users, and it is the most important one in forecasting process. Thus, a basic soccer knowledge should be required. The goodness of customization will determine system results.

Computing problems

The way of competition causes to take into account some special situations for computing attributes values. In this section we describe the main problems emerged and how they were fixed. Roughly speaking, these main problems concerns to initial matches in season.

Beginning of a new season: week 0

This problem is not hard, but as many others unavoidable, and a solution becomes essential. It happens when computing an attribute value related to league standings to forecast first week of a season. As any previous week has been played yet, there is not way to build a standing table.

When teams in current season remain in the same league as last, a trivial solution is to take into account positions and matches in last weeks of previous season. If the team played in a higher division than last season, it will be at the first position in the standing. Otherwise, if the team played in a lower division, it will be considered at last position.

Missing matches in attribute computation

Other problem, closely related to previous one, is when not enough previous matches are available to compute an attribute. Solution pass through taking lasts matches of last season as if they were in a continuous temporal line. This is not so simple, because of some teams were not playing at same division last season. Indeed, when playing in a lower or higher division, difficulty of division changes and matches cannot be compared into the same way. Therefore, we need to handle the situation of a team playing in a different division from current season division.

Other troubled situation where there are not enough matches for attribute computation is to compute results for directed matches between two teams because of there is only a few of such matches in the data source.

For these two related situations we offer two solutions. First is to compute attribute with a null value, but in this way we are giving a fake information to the system. We are setting that attribute is not true but, in fact, we have not information enought to determine it, so a better approach is required. Chosen solution is based on adjusting attribute's threshold. The value of this threshold is decreased proportionally to relation between number of required matches and number of available matches. Threshold γ is revised by

$$\gamma_{new} = \gamma_{old} \cdot \frac{\text{number of match results available}}{\text{number of match results needed}}$$

When number of required matches is too high and number of available matches is low, it looks like we are giving fake information to system again, but our experience shows that collateral effects of this approach are worthless compared to compute attributes with a null value.

Attribute selection vs expert system behaviour

In general terms, current base attribute set behavior forecast the most possible results of a match is quite good, in regular conditions. Even so, some experiments, in order to study attribute's behavior, have been developed.

Strict attributes

An attribute is *strict* when only a few objects can satisfy it, because of its threshold is too high. By working with sets of strict attributes, we can assure that they estimate the teams behavior better than other sets. Thus, with strict attributes, we will have very reliable estimates, but just only for very few matches, and non for most of others. In the other hand, using less strict attributes, system will produce less reliable estimations but for a big scope. So it is essential to find a balance between these two opposite situations: reliability of attribute set against number of matches without information. A good solution could be to build and use different attribute sets, ones more strict and others less. Thus, less strict attribute sets will be used when strict ones fail doing an estimation.

Trends towards the victory of the home team

It is a fact that, in soccer, it is more probable a victory from home team than away team. To deal with this, we offer two different approaches. First, modelling the teams behavior and second computing confidence values. For modeling teams behavior (attribute set customization) it is a good practice to use attributes with low exhaustive thresholds for

Factors	Correlation Degree	Used?
Associated to the classification in the league		
Team in the first classification level	medium/high	yes
Team in the last classification level	medium/high	yes
Difference between team's classifications	medium/high	yes
Team was in a different league last year	medium	no
Team socred a important number of goals (in the last matches)	medium/low	no
Associated to previous results of the team		
Number of consecutive won matches.	high	yes
Number of consecutive lost matches.	high	yes
Number of consecutive draws.	medium	yes
Number of non consecutive won matches in previous weeks.	high	yes
Number of non consecutive lost matches in previous weeks.	high	yes
Number of non consecutive draws in previous weeks.	medium/high	yes
Points collected in previous weeks.	medium/high	yes
Factors related with directed matches (incluidas previous years)		
Number of wins in previous directed matchs	medium/high	yes
number of losts in previous directed matchs	medium/high	yes
number of draws in previous directed matchs	medium/high	yes
Other Factors		
Number of red cards collected by the team's players.	low	no
Wheather the day and the city where the match took place	medium	no (hard to parametrize)
Motivation because of the fans support when playing as home team.	high	no (hard to parametrize, subjective)
Team hires a new coach.	high	no (only useful when new coach hired)
Some players of the team are selected for their National Team.	medium/Low	no (relevant for some nationalities)
Difference between team's budgets.	high	yes
One or more important team's players are injured.	medium	no (hard to automatically collect the data)
Cups collected in the lasts years.	low	no (only for a few of teams)

Table 1: Factors considered for selecting/building attributes

home team and more exigent threshold for attributes related to away team. Therefore, it will be easier for home team to satisfy an attribute than away team. It is possible to imitate this trend based on this approach.

Around 50% of played matches finish with victory of home team. This means that the attribute value, corresponding to matches result, will be 'home team victory' around 50% of objects from formal context. As consequence of the former, many rules from the inferred association rules will contain the attribute 'result = home team victory' within their conclusions. Thus when forecasting a match the system will infer, in most of cases, 'home team victory' as consequence of overestimation confidence value for this result. It is possible to avoid this effect easily, just applying a decreasing (reduction) factor over confidence for 'home team victory'. It is estimated by means of experiments.

Results

Following the process described above, an experiment was run for the Spanish premier soccer league from 2009-10. Attributes were selected according the experience of an expert, and contextual KB is computed (in Fig. 5 a KB fragment for *Málaga-Sevilla* match is shown). From this selection $\vdash_\exists$ is computed for each match in each week.

2009-10 season: Experiments with the system show forecasts of about 58.16% by a contextual selection based on the previous 38 matches of each team. Such a percentage of hits for a qualitative reasoning system may be considered as an acceptable result comparable with expectable results of experts (Goldstein & Gigerenzer 2009;

Attribute	Configurable parameters
1) Number of non consecutive won matches in previous weeks > threshold	<threshold> <Team> <Number of Matchs> <Matchs>
2) Number of non consecutive lost matches in previous weeks > threshold	<threshold> <Team> <Number of Matchs> <Matchs>
3) Number of non consecutive draws in previous weeks > threshold	<threshold> <Team> <Number of Matchs> <Matchs>
4) Points collected in previous matches> threshold	<threshold> <Team> <Number of Matchs> <Matchs>
5) Position in the classification based on previous matches> threshold	<threshold> <Team> <Number of Matchs> <Matchs>
6) Number of positions over the opponent in the classification based on previous matches> threshold	<threshold> <Team> <Number of Matchs> <Matchs>
7) Number of positions under the opponent in the classification based on previous matches> threshold	<threshold> <Team> <Number of Matchs> <Matchs>
8) Number of wins in previous directed matchs (included previos leagues) > threshold	<threshold> <Team> <Number of Matchs> <Matchs>
9) Number of losts in previous directed matchs (included previos leagues) > threshold	<threshold> <Team> <Number of Matchs> <Matchs>
10) Number of drawns in previous directed matchs (included previos leagues) > threshold	<threshold> <Number of Matchs> <Matchs>
11) Position in the classification > threshold	<threshold> <Team> <Matchs>
12) Number of positions over the opponent in the classification> threshold	<threshold> <Team> <Matchs>
13) Number of positions under the opponent in the classification> threshold	<threshold> <Team> <Matchs>
14) Number of consecutive won matches> threshold	<threshold> <Team> <Matchs>
15) Number of consecutive lost matches> threshold	<threshold> <Team> <Matchs>
16) Number of consecutive draws> threshold	<threshold> <Team> <Matchs>
17) Team's budget Y times bigger than opponent's budget (Y > threshold)	<threshold> <Team>
18) Team's budget Y times smaller than opponent's budget (Y > threshold)	<threshold> <Team>

Table 2: Attributes and parameters

```
....
14 < 2 > ID_2_T_0 ID_6_T_9 =[100%]=> < 2 > ID_1_T_16;
15 < 1 > ID_2_T_0 ID_0_T_16 =[100%]=> < 1 > ID_4_T_7 Team1_Wins;
16 < 2 > ID_3_T_2 ID_4_T_7 =[100%]=> < 2 > ID_0_T_16;
17 < 1 > ID_3_T_2 ID_5_T_7 =[100%]=> < 1 > Team1_Wins;
18 < 2 > ID_3_T_2 ID_0_T_16 =[100%]=> < 2 > ID_4_T_7;
19 < 1 > ID_3_T_2 Team2_Wins =[100%]=> < 1 > ID_4_T_7 ID_0_T_16;
20 < 2 > ID_4_T_7 ID_1_T_16 =[100%]=> < 2 > Team2_Wins;
21 < 1 > ID_5_T_7 ID_6_T_9 ID_1_T_16 =[100%]=> < 1 > Drawn;
22 < 1 > ID_5_T_7 ID_6_T_9 Drawn =[100%]=> < 1 > ID_1_T_16;
23 < 10 > ID_2_T_0 Team2_Wins =[90%]=> < 9 > ID_1_T_16;
24 < 8 > ID_2_T_0 ID_5_T_7 =[88%]=> < 7 > ID_1_T_16;
25 < 4 > ID_2_T_0 Drawn =[75%]=> < 3 > ID_1_T_16;
26 < 4 > ID_3_T_2 =[75%]=> < 3 > Team1_Wins;
....
```

Figure 5: KB fragment from Fig. 4

Andersson et al. 2003). Experiments with other contextual selections shows an increase in the number of hits by about 7% in the second half of the season. The reason is that data from the first half provides more recent information on teams and past matches.

2010-11 season: According to the idea commented above, we have evaluated the system in the second half of 2010-11 soccer season. A way to evaluate how good is this forecasting sistem is comparing number of successes in our pool with the most popular betting selections. This popular selections are collected from the most voted results for each match, published at state agency web that controls soccer pools. In Fig. 6 both results are compared. Our hits are in blue and popular ones in green and last seventeen weeks from 2010-11 season are represented. Note that Spanish soccer pools are over 15 matches.

Conclusions and Future Work

The challenge to detect emergent concepts for reasoning about complex systems represents an exciting researching field. Concepts with qualitative nature are extracted from data only considering partial features of complex system dy-

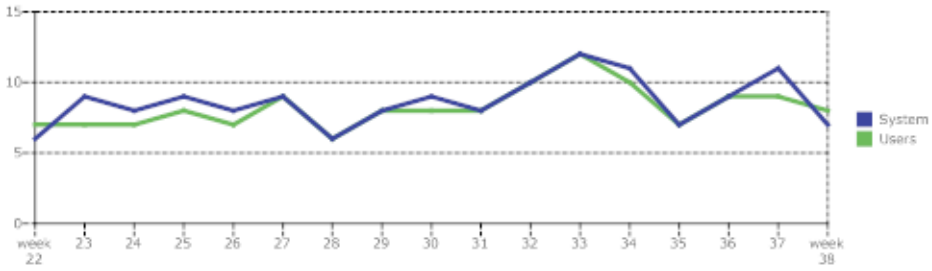


Figure 6: Correct predictions on the last 17 weeks of the season 2010-11

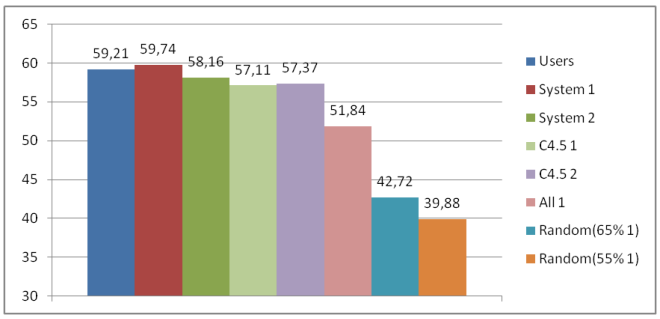


Figure 7: Comparative of correct predictions on the whole season 2010-11. Percentages

namics, a partial understanding of system. In this paper, FCA is applied to this aim with a specific application. The selection attribute problem based on FCA-based reasoning system for sport forecasting is analysed. In fact, the reasoning system is a computational logic model for bounded rationality. The model is concerned with association rule reasoning and it does not use -in its current form- more sophisticated probability tools (as for example (Min et al. 2008)). As is stated in (Goldstein & Gigerenzer 1996), the theory of probabilistic mental models assumes that inferences about unknown states of the world are based on probability cues (Brunswik 1955). It can say that confidence of association rules extracted from subcontexts play the role of probability cues.

Any statistical approaches have been taken into account, because of it was not the aim of this paper. Although a comparative study of our system against C4.5 classifier has been done. For this, two different attribute selections have been considered and used for both, C4.5 classifier and our system. The experiment is to forecast all matches (380) in season 2010-11. In order to stimate each match result, considering N (weeks) as timestamp, previous matches are used to build contextual selection (or trainining set in C4.5) from weeks $N-1$ to $N-19$ (190 objects). Fig. 7 shows the percentage of correct predictions for our system and C4.5 classifier, using both attribute selections. Other cols are also shown: 'user's most voted results', local team always win and two random generated. These 'random generated' cols were built assuming different weigths per result. It means, $< 1 : 55\%, X : 23\%, 2 : 22\% >$ and $< 1 : 65\%, X : 18\%, 2 : 17\% >$ were used, where $1, X, 2$ are the probabilities for forecasting a match with the result: local team wins, drawn and away team wins, respectively.

It is worth to note that, while classifier achieves highest performances (58,68%) when number of matches increase from 190 to 380, our system reaches this highest performance (59,74%) using only 190 instances. This conclusion is based on our system use some *fast and frugal* (Goldstein & Gigerenzer 1996) methods, and these are designed to achieve aceptable results using as less as possible resources.

The relationship of our proposal with Recognition Heuristics (Goldstein & Gigerenzer 2002) (roughly speaking, if one of the possibilities is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion) is not clear. We may assert that our model recognises trends in contexts. Trends (represented as association rules) can be considered as a kind of recognizing method, though. The system is based on bounded rationality models instead of statistic models, although in future hybrid models will be considered.

In the short term, we carry on extending our system in order to be able to combine the results of two or more attribute sets with different exigency level. Therefore the system will return only one result and more reliable. In the long term, we aim to extend the model in orfer to obtain a general system

to detect emergent concepts in Complex Systems

After some real betting experiments during current season (2010-2011) with one customized attribute set, we have observed another intriguing fact. If we take a look to number of successful predictions per week, we are able to distinguish some groups of consecutive weeks in which number of correct predictions is under or over the average. Recall that these predictions are the logical inferred results by one customized attribute set. This suggests that it could be possible to find another attribute set, with a different parameters customization, which it will accomplish the correct predictions of first attribute set. It means that when first attribute set produce bad forecasting, second should produce good ones, and vice versa. The reason of this is that each match there is not only one possible logical result. It means, when one of firsts teams of current ranking plays against one of lasts team, attending to ranking criteria, the logical result of this match would be that first one wins. But if we attend to others, like first team lost last week and second team won last 5 weeks, this results would be different. Future works pass through for finding these complementary attribute sets and detecting when their behaviors change during season in order to select the proper attribute set to forecast each week.

Finally, we are also analyzing how to finde a weight for matches which allows the system to work with matches from different divisions, simultaneously. Note that a winning match at first division will have a higher weight than a winning at second. This will be really useful at the beginning of season because of we need to compute attributes related to previous matches results and teams which are involved played at different divisions last season.

Acknowledgements

Supported by TIN2009-09492 project of Spanish Ministry of Science and Innovation, and *Excellence project* TIC-6064 of *Junta de Andalucía* cofinanced with FEDER founds.

References

Why Spain will win..., Engineering & Technology 5 June - 18 June 2010.

J. A. Alonso-Jiménez, G. A. Aranda-Corral, J. Borrego-Díaz, and M. M. Fernández-Lebrón, M. J. Hidalgo-Doblado, Extending Attribute Exploration by Means of Boolean Derivatives, Proc. 6th Int. Conf. Concept Lattices and Their Applications (CLA2008), pp. 121-132 (2008).

P. Andersson, M. Ekman, J. Edman, 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.

G. A. Aranda-Corral, J. Borrego-Díaz, Reconciling Knowledge in social tagging web services. Proc. 5th Int. Conf. Hybrid AI Systems (HAIS 2010), LNAI, vol. 6077. Springer-Verlag, Berlin, 383-390 (2010).

G. A. Aranda-Corral, J. Borrego-Díaz, J. Galán-Páez, Confidence-Based Reasoning with Local Temporal Formal Contexts. to appear in IWANN 2011, LNCS (2011).

G. A. Aranda-Corral, J. Borrego-Díaz, J. Galán-Páez, Bounded Rationality for Data Reasoning based on Formal Concept Analysis. To appear in DEXA Workshop DALI (2011).

W. Armstrong, Dependency structures of data base relationships. Proc. of IFIP Congress, Geneva, 580-583 (1974).

J.L. Balcázar, Redundancy, Deduction Schemes, and Minimum-Size Bases for Association Rules, Logical Methods in Computer Science 6(2):1-23 (2010).

E. Brunswik, Representative design and probabilistic theory in a functional psychology. Psychological Review, (62):193-217 (1955).

B. Ganter and R. Wille. Formal Concept Analysis - Mathematical Foundations. Springer, 1999.

J. C. Giarratano, G.D. Riley, Expert Systems: Principles and Programming. Brooks/Cole Publishing Co (2005).

D. G. Goldstein, G. Gigerenzer, Reasoning The Fast and Frugal Way: Models of Bounded Rationality, Psychological Review 103(4): 650-669 (1996).

D. G. Goldstein, G. Gigerenzer, Models of ecological rationality: the recognition heuristic, Psychological review, 109(1): 75-90 (2002).

D.G. Goldstein, G. Gigerenzer, Fast and frugal forecasting. International Journal of Forecasting, 25, 760-772 (2009).

Guigues, J.-L., Duquenne, V.: Familles minimales d' implications informatives resultant d'un tableau de donnees binaires. Math. Sci. Humaines 95, 5–18 (1986).

S. P. Imberman, B. Domanski, R. A. Orchard: Using Booleanized Data To Discover Better Relationships Between Metrics. Int. CMG Conference 1999: 530-539

B. Min, J. Kim, C. Choe, H. Eom, R. I. McKay, A compound framework for sports results prediction: A football case study. Know.-Based Syst. 21(7):551-562. 2008