Aalto-yliopisto
Teknillinen korkeakoulu

Aalto University

School of Science and Technology

Faculty of Information and Natural Sciences

Degree Programme of Computer Science and Engineering

Antti Rauhala

Pair Expression:
Re-expression Based Machine Learning Technique

Master's Thesis

Helsinki, April 30, 2010

Supervisor:	Professor Harri Lähdesmäki
Instructor:	Matti Haavikko, M.Sc.

Aalto University School of Science and Technology Faculty of Information and Natural Sciences Degree programme of Computer Science and Engineering		ABSTRACT OF THE MASTER'S THESIS
Author: Antti Rauhala		
Title: Pair Expression: Re-expression Driven Machine Learning Technique		
Number of pages: 110	Date: April 30, 2010	Language: English
Professorship: Information and Computer Science	Code: T-61	
Supervisor: Harri Lähdesmäki		
Instructor(s): Matti Haavikko		

Abstract:

This paper introduces a new technique for machine learning that is based on a brand new approach. Pair-expression attempts to find simpler and more dense expression for data so, that unknown variables becomes easier to predict and data is easier to compress. So in fact the technique is re-expression technique, but it has been designed for and it can be succesfully used for machine learning. Combined with naive bayesian predicting, it eliminates efficiently the bias resulting from naive assumption, and it can lead to even dramatic reduction in the error depending of the sample. As a result, naive bayesian no more functioned as a mere classifier, but as a predicter which provided (approximately) bialess probability estimates for unknown variables.

So as a difference to traditional machine learners, the technique attempst to optimize the information expression. In this problem setting, the aim is to find an optimal language L, that can be used for re-expressing the original data in form, where the redundancy between variables has been minimized

and as a consequence the regularities present in the data are captured in the language's structure. During language construction, surprisingly common variable pairs are re-expressed by introducing new expression variables. The redundancy introduced by the new expresison variables is eliminated with a special technique called 'variable reduction'.

Technique's properties were examined with a thought play, where the initial independence assumption is equated with classical analysis (where problems are divided into subproblems) and re-expression is equated with classical synthesis (where solution are formed from subsolutions). In the method the 'synthesis' is targeted against regular subsystems, which is considered to reduce approximation error with ideally small price of complexity and training error.

Technique's performance was evaluated by teaching it seven samples from 1995 Statlog study and by comparing results against 22 machine learners' public results. In testing pair-expression produced superior results for data, which consisted mostly of discrete variables, and it was first or second in four of seven samples, but with purely numeric samples the results were mediocre. The results were interpreted as extremely good and promising, especially because there is still lot to develop in actual algorithms. Especially the predicted variables could not be included for re-expression, because problems with prediction algorithm, which limited the learning ability. Based on experiences of this study, expression driven learning appears as very fruitful grounds for future research.

Keywords: Re-expresssion, machine learning, compression, re-expression driven learning, language learning, pair expression, statistical learning, naive Bayesian, knowledge representation

Aalto-yliopisto
Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietotekniikan tutkinto-ohjelma/koulutusohjelma

DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Antti Rauhala

Työn nimi: Pari-ilmaisu: Uudelleenilmaisuun perustuva koneoppimistekniikka

Sivumäärä: 110	Päiväys: 30.4.2010	Julkaisukieli: Englanti

Professuuri: Tietojenkäsittelyekniikka	Professuurikoodi: T-61

Työn valvoja: Harri Lähdesmäki

Työn ohjaaja(t): Matti Haavikko

Tiivistelmä:

Tämä paperi esittelee uuden tekniikan koneoppimiseen, joka perustuu uudelle lähestymistavalle. Tekniikassa pyritään hakemaan yksinkertaisempaa ja tiiviimpää ilmaisua datalle siten, että tuntemattomat muuttujat on helpompi selvittää ja että data pakkaantuu pienempään tilaan. Varsinaisesti tekniikka on siis tiedon uudelleenilmaisemistekniikka, mutta se on suunniteltu ja sitä voidaan soveltaa menestyksekkäästi koneoppimiseen. Käytettynä naiivin Bayesialaisen ennustajan kanssa, se eliminoi tehokkaasti naiivista oletuksesta johtuvaa systemaattista harhaa, ja voi johtaa jopa dramaattiseen ennustusvirheen pienenemiseen riippuen otteesta. Seurauksena naiivi Bayesialainen ei toiminut niinkään luokittajana, vaan ennustajana, joka tarjosi (likimain) harhattomia todennäköisyysestimaateja tuntemattomille muuttujille.

Erona perinteisiin koneoppimismenetelmiin tekniikalla pyritään siis optimoimaan tiedon ilmaisua. Ongelman asettelussa pyritään löytämään optimaalinen kieli L, jolla alkuperäisen datan voidaan ilmaista uudelleen muodossa, jossa esityksen muuttujien välinen redundanssi on minimoitu ja seurauksena tieto datan säännönmukaisuuksista tallentuu kielen rakenteeseen. Kieltä muodostettaessa alkuperäiseen tiedon esitykseen lisätään yksitellen uusia pari-ilmaisumuuttujia ilmaisemaan yllättävän yleisiä tilapareja. Ilmaisujen lisäyksien jälkeen käytetään muuttujien vähennystekniikkaa, jolla eliminoidaan ilmaisumuuttujan järjestelmään tuoma redundanssi.

Tekniikan ominaisuuksia tarkasteltiin ajatusleikillä, missä pohjaoletuksena tehty riippumattomuusoletus rinnastettiin klassiseen analyysiin (jossa ongelma jaetaan osaongelmiin) ja uudelleenilmaisu rinnastetaan klassiseen synteesiin (jossa ratkaisu kootaan osaratkaisuista). Menetelmässä 'synteesi' on kohdistettu säännöllisiin osajärjestelmiin, minkä katsottiin vähentävän approksimointivirhettä jopa ideaalisen pienellä monimutkaisuuden ja opetusvirheen hinnalla.

Tekniikan suorituskykyä arvioitiin opettamalla sille seitsemän otetta 1995 Statlog tutkimuksesta ja vertaamalla tuloksia 22 koneoppijan julkisia tuloksia vastaan. Testauksessa pari-ilmaisu tuotti ylivertaisia tuloksia datalle, joka koostui pääasiassa diskreeteistä muuttujista ollen ensimmäinen tai toinen neljälle otteelle, mutta puhtaasti numeerisilla otteilla tulokset olivat keskinkertaisia. Tulokset tulkittiin erittäin hyviksi ja lupaaviksi, erityisesti koska itse algoritmeissä on vielä paljon kehitettävää. Erityisesti on mainittava, että ennustattavia arvoja ei otettu uudelleenilmaisuun mukaan johtuen ongelmista ennustamisalgoritmin kanssa, mikä rajoitti oppimiskykyä. Perustuen tutkimuksessa saatuun kokemukseen, uudelleen-ilmaisuun pohjautuva oppiminen vaikuttaa erittäin lupaavalta alueelta lisätutkimukselle.

Asiasanat: Uudelleenilmaisu, koneoppiminen, pakkaus, uudelleenilmaisuun perustuva oppiminen, pari-ilmaisu, naiivi bayesialainen, kielen oppiminen, tilastollinen oppiminen, tiedon esittäminen

Contents

1Introduction

What is intelligence? This question has ever tempted the curiosity of human mind. What a wonder this gift of mind truly is; how unique and how powerful? What has the history of science proven for us, but that there seems to be no limit for human understanding. The wonders from the world, from the stars to the lightning to the mystery of life, even the greatest of the ancient secrets and mysteries have not escaped human understanding, but instead become trivia; everyday knowledge; food for the minds of our children. The ever deepening and improving view on the world reveals not only the incredible and enormous complexity that is in-build in this reality, but also the incredible ability of our limited minds to comprehend and understand something, that in its essence seems infinitely complex. The essence of this world is captured in words, shapes, memories and thoughts and in our minds it is understandable and it is predictable; and the way how we know the world guides ours hands and decisions in our daily lives. It is like a city in a bottle, the infinite complexity of the world trapped in the small space behind our eyes. Yet we manage in our everyday lives. Yet the great wonders of the world appear understandable for us, except for the mystery that we encounter when taking the very first steps of our lives; either on the surface of the water or on a glass of a mirror; or when we turn inside to wonder the very essence of ourselves.

Of course this essence of our minds that is intelligence is to still remain a mystery. This even when the Greek devised formal tool of formal logic to express understanding and to derive and validate conclusion; a tool that become almost like a living mind on the paper; a tool so powerful that it became almost a synonym for reason. This even, when advanced mathematical theory was introduced with even more powerful logic for reasoning in the form of statistics and probabilities. This even, when the computers emerged and so did the many solutions of artificial

intelligence, decision making and reasoning. This is even, when a wide scientific front emerged, each section carefully specializing on different fields like brain study, problem solving, machine learning, knowledge representation, planning/decision making, learning, language processing, motion/manipulation, perception, social behavior, creativity and general intelligence. Centuries have passed and so much has been sacrificed and given, yet the problem remains.

This paper is another small attempt in the wide field of artificial intelligence, and it concerns topics around machine learning and knowledge representation. Machine learning concerns the ability to learn from observations and to make predictions of unknown variables based on patterns recognized in the data. Knowledge representation concerns expression of knowledge; of many things, let them be variable states or patterns or many things and on how to use this information to derive further knowledge. Now, one may question, why the topics of *both* machine learning and knowledge representation are present in this paper? After all, these are often viewed as different fields of study; machine learning concerning patterns, while knowledge representation concerns expression of knowledge. Shouldn't they be separate? One may even say, that the only connecting link is on how knowledge representation tries to express the patterns machine learner tries to reveal.

Now, other could argue, that these topics are - in fact – strongly interconnected and even inseparable. There is an idea, that optimal or even meaningful expression of knowledge is heavily dependent of the expressed data and the patterns present in this data. Just consider the world we live in. In its essence, in microscopic level, it reveals its complexity, where even the smallest shapes consist of billions and billions of particles like a sea of small wheels ever living and turning. Still, on whatever level we view the world, let it be through telescope, naked eye or microscope, simple shapes and patterns form before our eyes and the world reveals itself as understandable. Galaxies, planets, lakes, trees, microscopic beings and molecules; our language as well as our thinking reflects the shapes and patterns of the reality

and this great regularity of the world - of how it repeats the same rules and shapes of the same behavior - is the sole reason why we can comprehend the world and why we can express knowledge and form memories. We cannot remember the ocean of restless pixels of our vision, but instead even the simplest details in our minds are in their essence patterns like shapes in our eyes that have continuity in time, in space and in reality. At this singularity of infinite yet super regular complexity, where even the smallest details are patterns of millions, the representation of knowledge has to be based on the observed regularities and the problems of knowledge representation and pattern learning becomes inseparable and one. Together they become the problem of approximation, the problem of modeling, the problem of dealing with complexity that is the grand problem of understanding, that how a system with million wheels can be reduced to a comprehensible and predictable form.

This new paraphrasing of the problem does not only combine pattern recognition, prediction and knowledge representation problems, but also the problem of compression. This should not be considered as a surprise, because all compression is based on re-expression of knowledge, and similarly all compression is based on developing better estimates for probabilities for pieces of knowledge, so that better codewords can be developed. Typically the only way to develop better estimates for probabilities is by recognizing patterns in the data and through this mean develop an expression or even a language for more efficient and compact expression. Fundamentally the compression problem is a probability prediction and a language problem, equivalent with machine learning and knowledge representation problems. The similarity can be demonstrated by comparing compression of the photographs with the way we understand the reality. Both of these are based on the heavy regularities present in the photographs and in the world. Good algorithm for photographs may provide compression of ten or hundred fold. The way we capture the complexity of the world inside our small heads seems to provide compression rate that seems in practice infinite.

At this point, let's get back to the artificial intelligence problem for a moment. We have paraphrased a number of problems to a single grand problem of understanding, but what is the relationship of this problem to the artificial intelligence field? Let's consider an agent acting in an environment resembling this reality. Ultimately the agent problem is a decision making problem; the agent holds purpose, which it seeks to advance, it receives information of the surrounding world and is entitled to actions that hold consequences on the environment. Greatly, the decision making problem is about understanding the consequences of the actions and understanding how these consequences further or undermine the raised purpose. In here, the decision making relies very heavily on the understanding and the knowledge extracted from the surrounding environment. Being able to predict the consequences of the actions is critical and to do so: knowledge of both details and patterns present in environment need to be utilized. Knowledge of patterns is needed for heuristics and to what we call common sense. Knowledge of details are needed for tasks like path finding, for familiarity with people, things and places and for most everyday activities. Often also the boundaries between details and patterns is fuzzy; one man's detail being another's pattern and vice versa; with even knowledge of details being impossible to form without utilizing the presence of regularities. Overall, it seems rather clear that decision making agent is very difficult to support with a machine learner or a database alone, but instead it would be best supported by further integrated mechanism that is the regularity based modeling; capable of both knowledge storing and prediction work; capable of dealing with all sorts of thing; big and small. In this context, this machinery's purpose would be no higher, but to provide the necessary understanding of the world for the artificial intelligence for it being able to make informed decisions.

Now, the problem of understanding can be approached through the problem of a language or expression. After all, all knowledge has to be expressed in some form; for men with words or letters, for apes with screams, for computers with bits. To

make the expression more powerful one has to give it structure and so a language emerges. To determine, that how the bits express the modeled complexity, one has to control not only syntax, but also the semantic meaning of expression, and the problem transforms into a form of knowledge representation. We seek to describe knowledge of some entity that we can call a system. When put to words, the problem of patterns and regularities becomes the problem of "What is the optimal language L for describing the system S?" (or even "What is the optimal language L for describing the system S to support purpose P?" that is more relevant for AI problem). The solutions for this problems serves both as a solution for knowledge representation problem and machine learning problem, because the language definition itself will reflect the patterns and regularities present in the system enabling learning and predicting, while the combination of the language definition and the re-expressed information captures the understanding and knowledge of the system. All this can be supported with statistical knowledge of the relationships between the concepts present in the language.

But for this paper, the attempt is not only to ask the grand question, but also to provide an answer for it; and this answer comes in the form of pair expression, which is in roots a tool for creating a language definition that can be used to re-express regular information in a heavily compressed form. The great trickery consist of using the language definition and statistical information to demonstrate understanding of the system in the way of making predictions for unknown variables and of acting as a powerful and well-performing solution for the machine learning problem.

So the research question for this paper is the question of understanding that is how to provide such an artificial construct, that it can captures the essence of examined system so that it can be used to re-express the information in a compressed form and it can be used for making predictions of unknown variables. It acts as an answer – even if as a limited one - for a number of problems that are the problems of learning, predicting, knowledge expression and compressing, and it stands in the middle

ground for these fields of study. It learns the regularities and patterns present in the complexity to understand it and form a simplified expression for the knowledge. Yet compared to the human mind, it is still a baby step and rather limited in its ability of regularity based modeling. Unlike the modeling of mind or many compression techniques for photographs, it is unlike to drop complexity by magnitudes. Instead, greatly because the compression/simplification is lossless, the rates are much more modest. Equally, it's ability to learn patterns and regularities are limited and instead similar to other machine learners, like neural networks or KNN. While it would appear to form classes or higher level concepts, this ability is also rather limited, even if it does perform very well when compared to a number of other machine learners as demonstrated in this study.

2Previous Study

The topic of this paper concerns quite wide field of study; having an emphasis on machine learning with the algorithm's background in compression and formal language learning, while having very special relationship to knowledge representation in the sense, that the algorithm's output can be interpreted as a such. While the topic concerns quite a number of areas, it does not build heavily upon existing machine learning, compression or language learning algorithms, even when it is possible to discover similarities with such. In a sense, pair expression takes some basic concepts from formal languages, introduces these concepts into the world of machine learning, introduces some semi-new ideas of system analysis/synthesis approach, variable re-expression/reduction mechanisms and finally builds some algorithms around these concepts based on quite basic information theory, statistics and university math.

While this paper is not relaying heavily on existing learning machines (except for naive bayesian), compression techniques (except basic theory found in the Shannon's famous paper) or formal language algorithms, the philosophy and quite lot thought work is based on various ideas present in the field of machine learning and computing science as a whole. Also comparison with similar ideas and techniques gives interesting context and a point of comparison for the introduced technique.

Before moving forward, the reader should know that there exist also implementation of pair expression algorithm for dealing with symbol sequences like plain text. This algorithm produced simplistic grammar, yet it was mainly designed fo compression and it provided similar if slightly weaker results for plain text as the very popular zip compression. To perform the compression, it analyzed the text to create a formal grammar that was by its expressive power weaker than regular expression. This

simple grammar learner / compression algorithm inspired the pair expressiont version introduced in this paper, that functions to solve machine learning problems instead solving patterns behind symbol sequences. This unpublished algorithm is utilized in following chapters to help comparing pair expression with various formal languages and their learners.

2.1Basics

This work relies heavily on the basic foundation of the information theory. The most important tools used when approaching problems are the basic ideas of entropy and information revealed in Shannon's famous paper “A Mathematical Theory for Communication”. While the basic approach is based on information theory, the information theory is again based on the wide foundation of probabilities and statistics, which are widely employed in this paper for both describing the binary variable systems and its properties and making predictions based on Bayes inference. Concept of 'naive assumption', the assumption of variable independence, is omitted from the context of the naive Bayes classifier, which is laid as an underlying assumption and corrected as dependencies are recognized. A major theme on this paper are considerations around Curse of Dimensionality coined by Richard Bellman, which emphasizes the problem of exponential growth in (state) complexity when amount of variables grows linearly. The complexity of problems is traditionally fought by a form of analysis, that is a process of dividing a complex problem into a number of smaller and less complex problems; an idea strongly present in Bellman's dynamic programming and an idea widely applied also in this paper. [1][2][8][9][10][11]

2.2Artificial Intelligence

The artificial intelligence is the study of intelligent agents. It has rather a long history, the topic being subject to philosophical debate from rather ancient times. It can be seen predeceased by the fields of formal reasoning and automation, before the era of computing and before emerging as a science of its own during 60s and 70s. It has been a subject of interests for computer scientists beginning from Turing from the very dawn of computing. It has seen its eras of progress and optimism, as well as its dark winters, which typically raised from recognition of incredibly challenging and seemingly unsolvable problems. [15][16][24]

The first golden era of artificial intelligence witnessed the emerging of the basic AI philosophical ground, symbolic reasoning, searching as reasoning as well as experiments with logic, simple neurons (perceptrons) and languages. This raise was brought to end by the limitations in the computing power and problems with exponentially growing complexity that is typical for artificial intelligence problems. Another boom followed during 80s with the raise of expert systems, knowledge revolution and back-propagation in neural networks, but exaggerated expectation were followed by disappointments and cuts in funding. Still meanwhile, quite a number of technical solutions brought by AI research have shown to be useful in a number of application fields, like the use of machine learners for expert systems, data mining, speech recognition and banking.

The field of artificial intelligence casts its shadow over most of topics covered in this paper. Artificial intelligence research has acted as a womb for fields like machine learning, knowledge representation and it is closely related to the sciences of formal languages and compression. Acting merely as an umbrella above the fields of more concrete and specialized sciences it provides interesting philosophical framework and proposes many questions that are relevant also for the topic of this paper. The early framework for the problem of artificial intelligence is presented in McCartney's 1969 paper “Some Philosophical Problems From the Standpoint of Artificial

Intelligence". This paper divides the problem of artificial intelligence into two parts, which are the heuristic part and the epistemological part, where the epistemological part captures the information of the world and the heuristic part uses the information to make necessary decisions to solve the problem setting. The questions raised in McCarthy's paper for epistemological part concern the kind of representation of the world, for its physical parts and for non-physical concepts like mathematics, goals and so forth, how observations are transferred into knowledge and how the system's knowledge is expressed. [16]

The problem of the epistemological part in its entirety is the artificial intelligence problem related to the subject of this thesis. A major part of the problem relates to the link between various concepts in the language and the observations and variables and states expressing the observed information. Another part relates to that how the concepts relate to the parts or shapes of reality. Pair expression learns patterns and shapes in the observations by creating expressions for abnormally common variable state combinations, which explains both how learning is done from observations and the relationship between learned concepts and observations. Now, considering that the observations are generated based on reality through some observation mechanism and the patterns, shapes and consistencies present in the world are expected to be translated as patterns, shapes and consistencies in the observation data. If the patterns present in the world become patterns in the data, the concept structure learned on the observations should ultimately reflect the structures present in the reality. This would explain the connection between expression concepts and the physical reality. After these considerations, the remaining problem is the encoding of knowledge as a sequence of symbols, which is somewhat trivia, because the representation/learning scheme is expression/language based. While the pair expression would appear to be somewhat complete answer to the epistemological problem in the sense that it provides some kind of answer for every aspect of the problem, it is not very powerful answer and its ability for pattern recognition is limited.

2.3 Machine Learning

The field of machine learning covers wide range of different models, algorithms and methods based on very different rationales and even very different views and definitions on learning. The different models draw their influence from various sources like quite basic mathematics, differential calculus, statistics, computation theory, information theory and from fields like evolutionary biology and neurology. Just for basic classification purposes a great number of different algorithms of different kinds can be identified. There are symbolic learners/classifiers associated with data mining as well as classical algorithms (K nearest neighbor, naive Bayes, linear discriminant, logistic regression) and more modern statistical algorithms (kernel density estimates) and numerous artificial neural networks (back-propagation, radial basis function) as well as many numerous other: Kohonen self organizing map, support vector machine, decision trees, bagged/boosted decision tree forests, different genetic algorithms, Bayesian networks and so forth. [4][6][7] [24]

It is needless to say, that the range of applications areas is even wider, but the various application areas can be also used to provide common metrics for evaluating the vastly different machine learning systems. The measurements of machine learning systems performance on different application areas can be used for making meaningful comparison between the numerous and vastly diverse solutions. While under the hood the comparison between learning systems may appear like comparing apples and oranges, on the application level they provide very similar or even the very same services of making predictions and revealing patterns and dependencies among different items or variables. While the many learning algorithms' performance vary depending of application and problem field, success in one application field

typically predicts success on other application fields and data sets. This consistency of success when facing highly varying random problems and data sets could be rightfully described as a generic learning ability. While it is far from an absolute metric, its approximate, that is the average ability to learn given a diverse data sets, has been used when comparing and ranking the performance of various systems in the past. Still even stricter metrics for performance can be obtained by analytical solutions that set boundaries for errors. For example Vapnik–Chervonenkis dimension is a formal technique for measuring the learning ability of statistical classification algorithms. [33]

There has been few comprehensive machine learner comparisons including StatLog: Comparison of Classification Algorithms on Large Real-World Problems in 1995 and An Empirical Comparison of Supervised Learning Algorithms in 2006. Statlog provided very interesting data on that day's non-commercial algorithm's performance, and it also attempted to find similarities and categorize algorithms based on their performance on the many samples. The results demonstrated e.g. the strength of statistical algorithms with credit standing problems, the strenght of classic machine learner's on segmentation problems and revealed presence of numerous similar patterns. What the 2006 study did show, was the strength of the aggregated 'democratic' learning machine, where the given answer to a problem is picked by a poll by a high number of weaker learning machines. In the study, this kind of 'democratic' boosted or bagged decision tree forests emerged as best performing solutions. These decision tree forests, which in the test situation could contain up to thousand individual decision trees, have been rather new development on the field. The idea was introduced in Michael Kearns unpublished manuscript of 1988 and the technique saw emergence in the 1990s. Of the more traditional systems, support vector machines, K-nearest neighbor and neural networks performed decently in the study and simple decision trees and naive Bayes classifiers became last. While the older study (StatLog) did not contain newer boosted/bagged decision

trees, it did measure performance of symbolic learners and statistic regression as an addition to other traditional methods. This older study did not appear to contradict the results of the newer study. [6][7]

The statistic/probabilistic solution in this paper is mainly a statistical algorithm and is closely related to naive Bayes classifier, in the sense that it attempts to improve naive Bayesian's results. Based on this knowledge, the algorithm can be expected to perform as well or better than naive Bayesian and have results that are similar to other statistical algorithms like Bayesian networks or logistic regression. Naive Bayesian in fact is not very well performing algorithm, but still it is worth noting that the main reason for the naive Bayes classifier's poor performance is the resulting bias, when the naive assumption does not really hold, and the solution seeks to fight this bias by modifying the information representation into a form, where the expressed variables are either exclusive or statistically independent.

2.4 Formal Languages

As this paper's topic is closely related to the expression of information, formal languages are naturally interesting. The study of formal languages came to life in 1956, when Chomsky formulated a mathematical model for grammar in connection of his study for natural languages. Shortly afterwards the connection between this formulation and programming languages was recognized, and the study of formal languages gave birth for syntax oriented compiling and techniques like compiler compilation. Also the connection between formal languages and various automata was recognized, and nowadays the languages and various automata are inseparable, like regular expression and state machine and stack automata. Also algorithms and Turing machine can be considered to be a pair of similar relationship. The

complexity of the language is closely related to the computational power of the machinery validating it. [19]

Formal languages and machine learning have an old yet strengtening relationship. According to the introduction of Jurafsky's and Martin's 2009 book “Speech and Language Processing”, the huge acceleration in language processing and speech recognition research since 1990 can be greatly attributed to the new emphasis on language learning that is supported by various machine learning techniques. The many versions of Markov Model has been utilized for ages to construct probabilistic state machines to approximate the structure of various languages. Nowadays support vector machines, maximum entropy techniques, multinomial logistic regressions and graphical bayesian models have become basic practice in computational linguistics. This development is supported by the emerge of various unsupervised statistical machine learning approaches. [18]

The current state of the art language learning techniques seems to be able to learn a subgroup of context free grammars. The abilities of various techniques were tested in the 2004 Omphalos Context Grammar learning competition. The winning algorithm is described in Alexander Clark's 2005 paper “Learning deterministic context free grammars: The Omphalos competition.”, which remarked the competition being perceived as extremely challenging. The algorithm used mutual information to construct the grammar, but was otherwise quite different compared to pair expression. It assumed text to follow context free grammar with a limiting non-terminally separated (NTS) property, and it was interested of surprisingly common symbols surrounding specific strings instead of surprising high co-occurrence of symbols or variable states as in pair expression. While Clark's algorithm provided the best performance in the competition, it is remarked in Clark's paper that the algorithm is able to learn only a subset of context free languages. [22][23]

When considering the significance of formal languages in relationship to pair expression; two connections between these things can be recognized. One connection

is the connection between formal languages and the machine learners in general. Another connection is the background of pair expression mechanism in the formal languages. In fact, the first 'version' of pair expression was designed for the purpose of generating a formal grammar and for functioning as a simplistic compression algorithm and for this purpose it did work. While the focus of the mechanism has changed from symbol sequence to binary variable system, quite many characteristics of formal languages were carried to this new re-expression / machine learner / compression algorithm. Still as a result of this metamorphosis, the pair expression language got loaded with semantic meaning in the sense, that its expressions carried information of the examined system. Despite application domain differences, the interpretation of pair expression as a formal language is interesting and somewhat revealing. Original pair expression was able to express very limited subset of regular expression; and its automata equivalent had been rather limited non-deterministic state machine. These 'limitations' in formal languages world still don't appear to be a handicap in the machine learning world, where the pair expressions perform well. One interpretation would be that neither other machine learners hold expressiveness of e.g. context-free grammar or computational power of e.g. stack automates, and they cannot have, because the data they are working on is not ordered/organized. Having organized data (e.g. binary variables set in sequential order or organized as a matrix) would perhaps allow machine learners with greater expressiveness that could be interpreted as more powerful automates. E.g. contexfulness could provide means for recognizing shapes (shape borders form NTS context) in a color matrix and the mindplay tempts interpretations of other language learner techniques as machine learners. If the data were ordered, could e.g. Clark's algorithm be generalized as a machine learner?

2.5 Compression

As people that are familiar with information theory know, compression is based on ability to assign accurate probabilities for pieces of knowledge (to get optimal codewords), making it very closely related to predicting. Naturally, compression is also related to re-expression in the sense that compression is re-expression; with an emphasis of reducing the re-expressed/encoded information size. In a away, both predicting and compression are based on the assumption that examined system is somewhat predictable and regular. This also helps to understand why there is a strong connection between machine learning and compression to the point that machine learning solutions can be combined with simple compression algorithms for very impressive results. [1]

While machine learning techniques can be used for compressing information, the compression field itself has developed algorithms for rather generic lossless data compressions as well as specialized algorithms for specific purposes like audio and video compression. The main data compression algorithms are based on finding optimal codewords for symbols in the data and eliminating recurring sequences of symbols or simply words. Examples of codeword assignment methods includes various Huffman encoding schemas and arithmetic coding. Lossy audio and video compression techniques typically transform the data into waveform and then eliminates components that hold only marginal effect on the outcome. Still, the very traditional and common means for compression hold only little relevance for the topic. [8][26][27]

The combinations of machine learners and compression algorithms are somewhat more interesting. Machine learner is feed with some parameters, like previous bytes or words, and it is simply used to assign probabilities for the upcoming symbol's alternatives which are then transformed to code words. In the simplest scenarios, the predictions can be based on plain statistics as it is in PAQ compression algorithm, which achieved the Hutter prize (awarded for compressing 100 MB text corpus of human knowledge) baseline in August 2006. As a curiosity, newest versions of PAQ

compression uses artificial neural network to combine predictions from various statistics. [22][23]

Similarly machine learners, which internals can be used for compression as such, are very interesting. An example of such is the Kohonen Self-Organizing Map, which neuron map 'stretches' to form an approximation of the trained variable space. Symbolic FP-tree, which is used to reveal association rules and which has background in data-mining, is similarly able to both make predictions and compress information, all thought as a difference to SOM the compression is lossless. Compression interpretations can be done for the internals of other machine learners as well. [25][26][28]

Even more interesting are the compression techniques, that assume the data to follow language structure and will compress the text based on a grammar generated from the text. An example of this is the language compression program SNPR introduced in J. Gerard Wolff's 1982 paper “Language Acquisition, Data Compression and Generalization”. SNPR's output actually resembles heavily pair expression's symbol stream version's output, except that it apparently uses simple statistical rules instead of information theory for expression forming, all though it seems otherwise more sophisticated and more powerful. The main difference is that after picking out common symbol sequences (what pair expression also does), SNPR's also seeks to find shared contexts within the sequences to construct a context-free grammar. [21]

Also methods that construct probabilistic automates can be interpreted as a language-driven compressors. E.g. algorithm that reconstructs a Markov model can be seen language driven compressor; as the results that is the probabilistic finite state machine can be interpreted as a regular expression kind of formal language. Here again there are similarities with pair expression, which constructs a grammar that is similar yet weaker than regular expression. Still, one major difference between Markov model and pair expression is that the pair expression can be generalized also for unsupervised and supervised problem settings of the machine learners.[19]

2.6 Knowledge Representation

Knowledge representation has important role for this paper, solely for the reason that the pair expression can be interpreted as a method of expressing knowledge as such; even if its abilities in expressing knowledge and patterns are somewhat limited. The field of knowledge representation studies expression of knowledge so that reasoning can be done, which basically means either validating or deriving new knowledge from the expressed knowledge base. Davis Shrobe and Szolovits article “What Is a Knowledge Representation?” provides five different aspects for knowledge representation, of which one concerns representation as 'replacement' of real world (author's comment: or raw data), a way of formulating problems, thinking and reasoning in the world and where remaining aspects consider the practical performance aspects and the human ability to understand the expression. [29]

In a sense, the McCartney's article of AI philosophy (discussed before) concerns quite wholesomely this topic and the epistemological problem of the paper is very closely related to the problem of knowledge representation. The main difference between the problems is that epistemological problem includes also the problem of machine learning (of how to learn from observations). For KR, while classes, patterns and rules function as targets of description, the aim is not to learn them from the raw data, but rather devise means to describe known instances, classes, data and regularities. One may even say that knowledge representation attempts to describe things that machine learners attempt to learn. Still, one can recognize a mismatch in the sense that the knowledge representation languages are able to describe far more powerful concepts and entities than any of the existing machines learners are able to ever learn. This mismatch is reflected by how the research around semantic networks

and knowledge systems concern problems that are from rather different world compared to the problems relevant for this paper. [16]

From the pair expression point of view, the most interesting question is the expressive power of various notations, e.g. what are the expression power differences between hierarchical and network models, classification and description models, and that what kind of information requires symbolism or recursiveness. This would provide a point of comparison for the expression and help in setting new and interesting goal for future research. According to Nebel's 2000 article “On the Compilability and Expressive Power of Propositional Planning Formalisms”, there is a wide consensus around the expressive power of various language features, but uncertainty on how to measure the expression power in a formal way. Currently still, the one commonly accepted criteria for comparison seems to be the conciseness of translating/compiling expression from one language to another, even if this compilation is not always computationally possible. For example, if some expression in language A translates into infinite series of language B expressions, one may conclude that language A is stronger at least in this one respect (author's note: this approach would indicate, that the expressive power of knowledge representation is related to its ability to compress information and therefore to its ability to express regularities). [30]

It is apparent, that there is a number of formal tools for measuring the expressive power for a knowledge representation, even if there is uncertainty and the debate continues. Such formal tools are presented not only in Nebel's article, but also in 1996 paper from Cadoli et. al. and in Gogic et. al. 1995. Still, formal study of the pair expression's expressive power would apparently be a major task and perhaps deserve a paper of its own. However, such study would likely show the expressive power of pair expression to be rather limited. Still in the paper's solution's defense, the 'standard' problem setting as well as common corpuses for machine learning set

their own strong limitations on how powerful patterns can be learned. Some kind of organization (sequential or matrix) or desription of the variable relations (e.g.