

Pattern recognition through perceptually important points in financial time series

G. Zaib, U. Ahmed & A. Ali

NUST Institute of Information Technology,

National University of Sciences and Technology (NUST), Pakistan

Abstract

Technical Analysis is a financial risk management practice that has been in use since the advent of the stock market and Pattern Recognition is an indivisible part of it. There has been a lot of research into pattern recognition in time series. Existing pattern recognition techniques lack dynamic extensibility. They do not provide any interfaces in order to include new patterns for recognition dynamically. This limits the operability of these techniques to a particular domain. This research devises a new technique for domain independent pattern recognition while giving sufficient speed and accuracy. This enables it to be used by critical Decision Support Systems for time series of different domains. The system emulates the human visual cognition process by implementing the concept of Perceptually Important Points Identification (PIPI). Perceptually Important Points (PIP) represents the minimal set of data points which are necessary to form a pattern. For dynamic inclusion of patterns a Pattern Definition Language (PDL) has been conceptualized for defining patterns in time series by using a declarative programming paradigm. This also results in domain independent pattern recognition without needing any modification.

Keywords: stock pattern mining, pattern recognition, financial time series, soft computing, declarative programming, time series analysis, perceptual modelling, computer vision, non-linear pattern recognition, modelling human cognition.

1 Introduction

Historic time series data is one of the most precious assets of modern business and scientific research entities. Modern decision support systems (DSS) are expected to execute important policies based upon patterns occurring in time series. This Makes pattern recognition and integral part of any future DSS.



Pattern recognition is primarily a human cognitive faculty. Depending upon an individual's ability faces severe limitations. Its expertise vary among individuals, thus it cannot provide basis for the decisions involving a high degree of risk. Secondly scanning huge amounts of data to for a single pattern or even recursively scanning it for multiple patterns is a humanely impossible task [1]. Still human cognitive process is indispensable as all the small non linearities and variation in patterns cannot be captured through traditional parametric approaches. So we develop a system which utilizes the strengths of human cognitive process while removing its above discussed weaknesses.

Existing work provides hard coded approaches that limit the applicability of the system in two aspects. Firstly it is hard to include new patterns in the knowledge base of the system as it needs retraining and redevelopment. Secondly this result in domain-specific pattern recognition techniques limited to a few patterns defined at design time. This makes the existing solutions fixed and inextensible. Extensibility requires the capability of dynamic inclusion of patterns, to be present in pattern recognition system on the one hand, and domain independence on other.

We develop a system that closely imitates human faculty of pattern recognition and is called Perceptually Important Point Identification (PIPI).

Briefly, it first divides the time series being scanned in to segments. Once segmentation is completed the whole system explores the time series to discover PIPs. Once pips in a specific formation are found, system classifies that segment of time series in which PIPs are found as making the desired pattern. To extend the system the user simply has to define the pattern in our specified PDL (Pattern Definition Language). The system would then be ready to start looking for that new pattern in large volumes of Time Series.

Another advantage of PIPI is that the system becomes capable of providing accuracy measurements of discovered patterns. This helps refining the pattern definition and provides a set of patterns with varying degree of closeness with the actual pattern. The PIPI is useful in such systems where we need fine tuned pattern recognition and we want to start with a few patterns in the beginning and add more patterns to it as the pattern get discovered.

Rest of the paper is structured as follows. First the status of current research in the field of pattern recognition is explained. A PIP (Perceptually Important Point) will be defined; also the advantages gained by using perceptually important points instead of traditional parametric techniques are given after that. Then whole PIPI is explained along with an example of stock time series. The structure of PDL and the architecture of its parser and interpreter will be given after that. The results of two performance gauging experiments done on PIPI are given and in the end there is conclusion.

2 Previous work

There has been a lot of research done in pattern recognition. All kinds of techniques from traditional parametric techniques such as neural networks to novel pattern matching techniques have been put forward.



Ge [8] has used hierarchical pattern analysis for pattern recognition in time series. Hierarchical pattern analysis means that the whole time series is divided into a fixed sized and number of segments, then those segments are compared one by one with those present already in the database, proceeding to next segment only if previous segment matches. The problems with this technique are a) a pattern must be seen at least once to be recognized, b) The user can not himself add patterns to it.

Fu [7] applied self organizing maps to the process of pattern discovery for making neural network training for pattern discovery efficient and accurate. Self-organizing maps are also called radial bias networks. The main problem with this technique is that it is not domain independent or extensible because the system recognizes only those patterns for which it has been trained.

Singh [5] has presented his technique of SNN (Single nearest neighbor) for pattern detection in long memory systems. This technique works by comparing the time series segment (in which the pattern is to be discovered) with the previous time series, and if it finds the segment with the same properties it classifies the segment to the same category. This approach is extremely slow and further becomes slower as the time series or size of the segment becomes larger.

Pattern matching utilizes distance measure by storing the Euclidian distance between fixed points, and then compares that distance with the previous time series to locate a pattern. However, there is no way a pattern can be detected if it has not occurred before, and it also becomes slow like SNN technique of Sameer Singh and there is no way a user can add his own patterns to the system to make it extensible [6].

Most of the research has been mainly focused on two attributes of pattern recognition i.e. efficiency and accuracy. There is still a need for a pattern recognition technique which should be able to work across domains, should be extensible enough, meaning that once it has been trained to recognize certain patterns at development time, it can be extended later to add more patterns. The features of extensibility and domain independence should not be achieved at the cost of efficiency or accuracy i.e. it should be efficient enough to be used in interactive system, should be accurate enough so that people could rely on it for critical decisions.

3 PIPI Perceptually Important Point Identification

3.1 PIP (Perceptually Important Points) definition

Time series are made up of infinite number of data points if we consider a pattern on a time series it will be subset of the points of making the whole time series, but even all those points are not required for pattern recognition. If we concentrate on even a smaller set of point we are still able to recognize pattern. These points, which are necessary for pattern recognition, are termed in this paper as Perceptually Important points or PIPS.

Mathematically PIPs can be represented as if we take time series as a set of data points X , then the pattern will be a subset of the set to time series data



points X i.e. $Y \subseteq X$ now perceptually important points will be a even smaller minimal sub set from which set Y can be derived, say if Z in the set of perceptually important points then $Z \subseteq Y$.

3.2 PIPi: emulating human cognitive process

Tarr [4] explains the mechanism of human recognition of patterns as a process requiring continuous perceptual analysis. Which says that when we observe something, we concentrate on increasingly detailed characteristics of the objects called precepts. For example, when we are observing an apple we will first take sensory account of the shape to recognize it is an apple, then we will go in to further detail of its 3d properties and color to see whether it's a painting or a real apple, then further we go in to the details to try to recognize the apple's family.

PIPi closely follows this human method of pattern recognition; it works by recognizing precepts know as "Perceptually Important Points", in given time series.

Before the recognition process starts, the time series is divided in to segments. This segmentation is done based on the maximum and minimum values a time series can take and a segmentation factor which will tell the size of segment.

After segmentation, it tries to figure out a pattern by step-by-step comparison. For this it first finds the foundation PIPS, then goes to the stage where the upper layer PIPs are found.

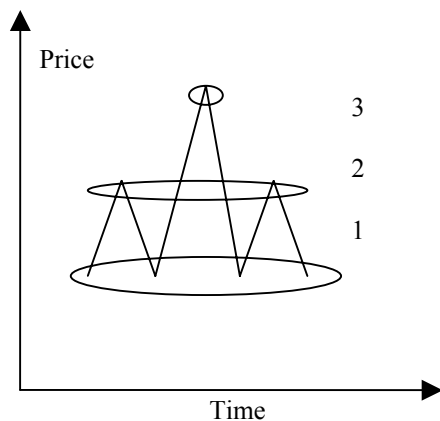


Figure 1: Step-by-step PIPi.

3.3 Challenges requiring the use of PIPS

- **Domain Independent:** The pattern recognition should work across domains, the technique which should be able to ignore differences in the properties of time series. It should be flexible to work across domains



- **Time Unit Independent:** Time series can be on varying time scales within and across domains so there should be such a technique which is able to work no matter the time scale is hours, years or any thing in between as long as it is constant within a pattern.
- **Highly Non-Linear Patterns:** Most of the patterns cannot be defined strictly in terms of shape. Small non-linearities will always exist in time and value. For example consider head and shoulder pattern of financial domain shown in Figure 2. Describes as $A=C=E=G$; $B=F$; B in mid of AC , F in mid of EG ; D in mid of BF . However it is not necessary that all points requiring equality in shoulders are always exactly equals nor it is necessary that the head lies exactly in middle of two shoulders so the technique used for pattern recognition should be flexible enough to cater for such non linearities.
- **Varying scales between domains:** If we talk about such a pattern recognition algorithm which would be able to work across domains then another challenge which must be overcome is varying scales of y values across and even within domains. For example stocks of Microsoft vary between 50\$ and 150\$ where as those of CNN vary between 5\$ and 10\$ but when we switch within domains the problem becomes even more acute as we have to cater for even more wide value swings. Consider ECG data which changes by .001 as compared to stock prices where change is usually several places before decimal.
- **Missing Data:** Most of the time several data values between 2 ranges of time series are missing. Thus we must also implement some method to overcome the problem of missing data.
- **Need for a smoothing function:** Time series having a lot of small non linearities confuse the traditional techniques, so the data is passed thorough filters in order to smooth it.

3.4 Advantages of using PIPI

- **Time Unit Independence:** When using PIPS we get independent from time unit, it does not matter what unit of time each data point of time series enclose as long as all the PIPs in the pattern belong to same time unit.
- **No need of a smoothing function:** When using PIPs there is no need of a smoothing function, as highly noisy data does not affect accuracy. This is because we do not need to feed whole time series to the technique. Only we need the data related to PIPs.
- **Missing data do not effect accuracy:** We do not need any filter to fill in missing data as long as the data points that will form the PIPs of the pattern exist, on these PIPs will be sufficient to do the pattern recognition.
- **Pattern Definition Feasible:** Further more when using PIPs, pattern definition is feasible as we only need some way to represent the PIPs in



the time series. It also represents their relative position in the pattern in order to make our algorithm domain independent and evolvable.

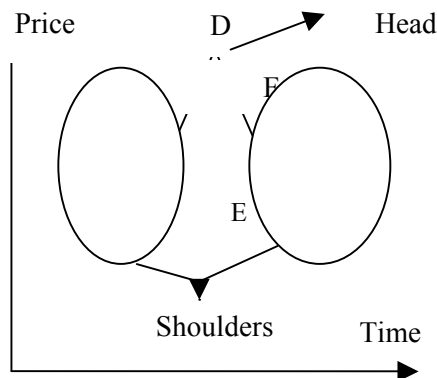


Figure 2: Head and Shoulder.

3.5 PIP identification algorithm

To explain the PIP identification algorithm this paper will build upon the example of head and shoulder pattern in Figure 3

- Find Minimum and Maximum value of time series for this time series. For this time series min and max values are 2.5 and 20.5 respectively.
- Divide whole time series in to segments with a segmentation factor. The segmentation factor is fixed for time series belonging to each domain for example it is 1.5 % for stock data. However it can too be varied to increase the accuracy of recognition algorithm for each domain. The density of points lower as we go up, so the band size must increase in order to allow for lowering density. For the sample time series the number of bands will be 24.
- Start PIP Identification Process once the time series has been segmented. For this we first; arrange the segments in order by their volume, because almost all the pattern starts from the segment having high density of data points. In this sample series this segment is the segment 1. The point indicated in Figure 4 are in that segment.
- Find Neighboring Pips that lie in the bottom layer, these are marked with red arrows, check if they satisfy the pattern i.e. in same band number is greater than four then we move on to upper layer so that the shoulder should be completed, now we will find

the points in the upper layer at the mean of point already found, but not at exact mean, give or take a single time unit thus point marked in white in Figure 6 would be found, this will result in recognition of two shoulders and then on the mean of these two we will find another point for the head which is marked in dashed in Figure 6.

- Remember the matching points found, if they become equal to a specified number then the pattern is successfully detected e.g. for Head and Shoulder Pattern there must exist at least eight points.

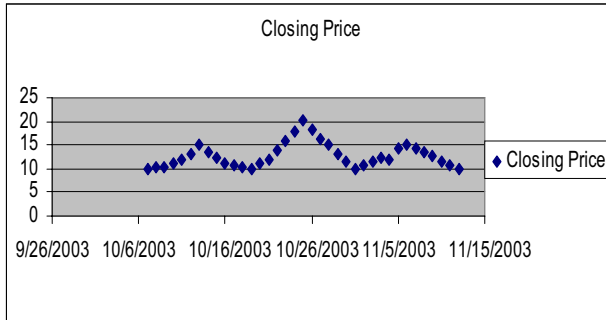


Figure 3: Sample.

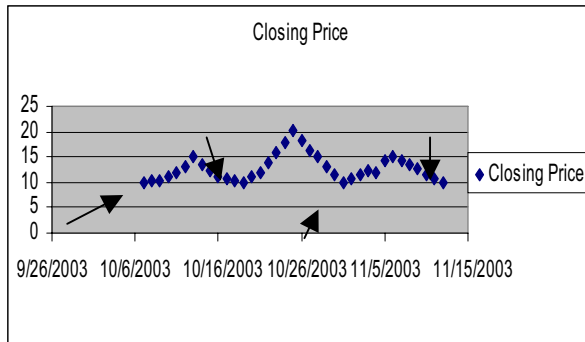


Figure 4: First layer PIPs.

4 Pattern Definition Language

4.1 PDL requirements

PDL (Pattern Definition Language) makes the system extensible, so it should be capable of modeling different aspects of the pattern. Each PIPs has two properties, the value on x-axis which is always some unit of time and the value

on y-axis which is different for each time series. To define the relation ship between PIPs we will also take:

- The distance of one PIP from another PIP in time.
- The distance of one PIP from another PIP in value.

One of the strengths of human cognition is its ability to approximate patterns. Approximation means that humans have the capability to ignore small difference and take a look at the bigger picture to recognize a pattern. Figure 5 shows several time series samples, all are head and shoulder patterns, though they show large difference in the slope of each line and also the length of each line. So it is clear that though we can define the relation ships between PIPs to make a pattern, but they cannot be strictly defined as there must be some variation involved.

So the PDL designed should be capable of:

- Defining the actual change in value that must be present between one PIP and the next.
- The variation that can take place in the value change, i.e. the number of value units around the original where a PIP can exist if it's not found at the original change position.
- Specify the next time interval where the PIP will be found.
- As it is quite possible that the next PIP is not exactly at the specified time interval so we must also specify a variation for time.
- The variation or interval may be defined with respect to any previous point found. For example if the last point we found is "n". A requirement comes to find point that is 3 time units a head of point n-3 i.e. a point found three points back. So the PDL should be capable of doing this.
- Moreover the variation can be till the end of time series so there must be some way to specify the end of time series along X axis or Y axis.

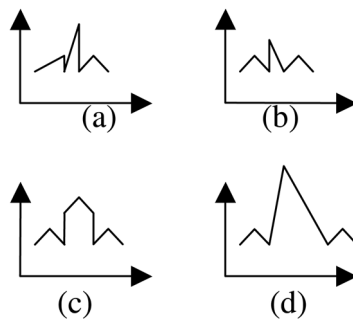


Figure 5: Variation in head and shoulder.

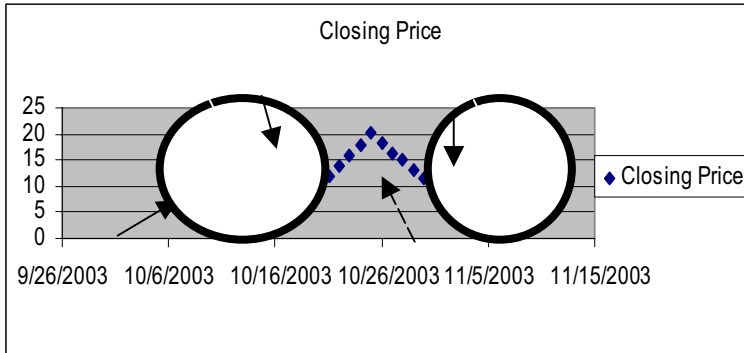


Figure 6: All PIPS in pattern are found.

4.2 PDL structure

The PDL should be such that it is easy to write and also can be parsed accurately and efficiently. Let's review what should be the structures present in PDL in order to define pattern

- Value Interval from one pip to next.
- Variation in value interval.
- Time Interval from one pip to next.
- Variation in time interval.
- References to previous point or end of time series.

Table 1 shows the different key symbols present in the PDL:

Table 1: PDL Key Symbols.

Symbol	Usage
N	Specify variation to the end of time series
+, -	Define the direction of finding next PIP: +: Move forward in time or move higher in value -: Move previous in time or lower in value
value	Any value written between indicate the reference to any previous point i.e. 2 means move to point found 2 points back.
/	The line separator used in PDL, just like the use of “;”.
+	Used to show the start of a pattern i.e. to indicate that system should pick up the first point to start looking for new pattern



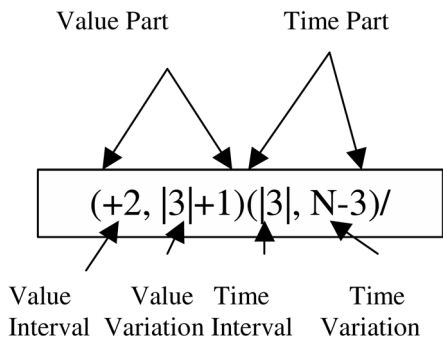


Figure 7: PDL Sample.

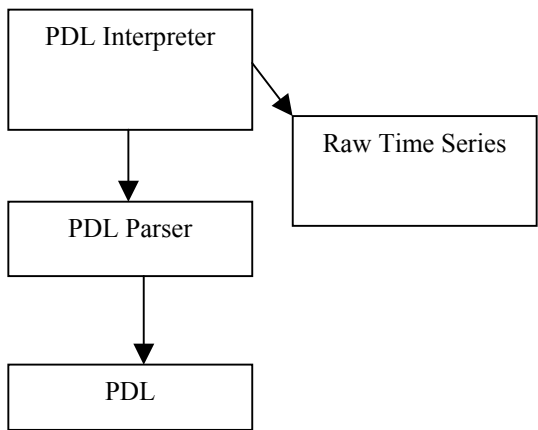


Figure 8: PDL Parsing.

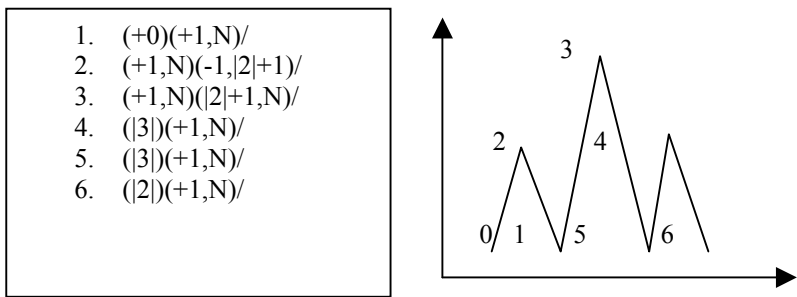


Figure 9: PDL: Example.



A sample PIP definition is given in Figure 7. The sample line commands the system for look for a PIP with following properties

- Two value units above current PIP and search for it in the reverse direction till the value unit that is 1 point greater then that of PIP found 3 PIPs ago.
- At the same time unit at which the a PIP found 3 PIPs ago was and the variation could be till three unit back from the last time unit in time series.

4.3 PDL example

As an example I will define the head and shoulder pattern given in Figure 9 along with commands.

4.4 Structure of PDL Parser and Interpreter

The architecture that results from the research, responsible for embedding PDL in to PIPI, is given in Figure 8.

When a pattern is in the process of being detected following activities are done by each part

- PDL defines the pattern.
- PDL parser reads the PDL and translates the symbols for the PDL interpreter. The parser will tell the interpreter that it should look for a PIP between this time unit to this time unit and also from this value unit to this value unit.
- PDL interpreter gets the instruction from the parser and try to find the PIP at the specified location in raw time series.

An advantage we get out of this architecture is high modularization, as time series searching code is present in the interpreter and PDL parsing code in parser. Thus if we want to modify the PDL or want to extend PDL, we will simply have to rewrite parser and our system will be updated.

5 Conclusion

PIPI is a simple yet powerful technique to recognize patterns. It improves over previous work in this area in the following aspects

1. It provides a high level abstraction of time series patterns, rather than treating pattern discovery process as a black box – something that the user may be uncomfortable with.
2. Its specification is close to how human keep patterns in memory, where as its working closely imitates the human way of pattern recognition.
3. PDL let us define patterns based upon the relationships between perceptually important points. It makes the system extensible in two aspects. Enabling dynamic inclusion of pattern and also the technique lends it self to be used across the domains without any modification.



A PIPI prototype implementation exhibits very impressive resource utilization during the detection process of the pattern and also shows highly optimized time series search. The experimental data was of year 1995 for share prices of CNN. Besides we also used a hypothetical data of 87 years for stress tests. A graphical user interface is the future direction of this. This shall enable visual usage of PDL. It will help the user to define a pattern by marking the segment of time series in which the pattern occurs.

References

- [1] Hedge, Human Information Processing – Pattern Recognition, Decisions And Attention, Cornell University 1999.
- [2] Frietman, XCG-NEUROPC: Senter R & D program, Faculty of Applied Physics, Delft University of Technology, Delft, Netherlands 1994.
- [3] Identifying Seismicity Patterns Leading Flank Eruptions At Mt. Etna Volcano During 1981-1996, Center for Theoretical Physics, Laboratory for Nuclear Sciences and Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 1999.
- [4] Tarr, Visual Pattern Recognition, Brown University 1996.
- [5] Singh, A Long Memory Pattern Modelling And Recognition System For Financial Time-Series Forecasting, University of Exeter Department of Computer Science Prince of Wales Road Exeter 2001.
- [6] Pratt LOCATING PATTERNS IN DISCRETE TIME-SERIES Department of Computer Science and Engineering College of Engineering University of South Florida 2001.
- [7] Fu, Pattern Discovery From Stock Time Series Using Self-Organizing Maps, KDD2001 Workshop on Temporal Data Mining.
- [8] Ge, Pattern Matching in Financial Time Series Data, third conference on Knowledge Discovery in Databases and Data Mining 2000.

