**1. Lý do chọn đề tài**

**Ngày nay, khi xã hội ngày càng phát triển thì lượng**

**thông tin càng tăng lên với tốc độ bùng nô. Lượng dữ liệu**

**khổng lồ ấy là một nguồn tài nguyên vô giá nếu như**

**chúng ta biết cách phát hiện và khai thác những thông tin**

**hữu ích có trong đó. Như vậy vấn đề đặt ra với dữ liệu của**

**chúng ta là việc lưu trữ và khai thác chúng. Các phương,**

**pháp khai thác dữ liệu truyền thống ngày càng không đáp**

**ứng được nhu cầu thực tế. Một khuynh hướng kỹ thuật**

**mới ra đời đó là Kỹ thuật Khai phá dữ liệu và khám phá trí**

**thức (Knownledge Discovery and Data mining - KDD).**

**Công nghệ khai phá dữ liệu ra đời đã cho phép ta khai**

**thác được những trì thức hữu dụng bằng việc trích xuất**

**những thông tin có mối quan hệ hoặc mối tương quan nhất**

**định từ một kho dữ liệu lớn (cực lớn) mà bình thường**

**không thể nhận điện được từ đó giải quyết các bài toán tìm**

**kiếm, dự báo các xu thế, các hành vi trong tương lai, và**

**nhiều tính năng thông minh khác. Ngày nay, các công**

**nghệ data mining được ứng dụng rộng rãi trong hầu hết**

**các lĩnh vực: phân tích dữ liệu, dự báo,**

**Một trong những vấn đề quan trọng nhất trong lĩnh**

**vực tài chíhhiện đà ôm kiếm**

**những cách thức hiệu quả để tóm \_ tắt và hình dung dữ**

**thị trường chứng khoánđể cung cấp**

**cho các cá nhân hoặc tổ chức những thông tin hữu ích về**

**các hành vi thị trường hỗ trợ việc ra các quyết định đầu tư.**

**Số lượng lớndữ liệu có giá trị được tạo ra bởi thị**

**trường chứng khoán đã thu hút được các nhà nghiên cứu**

**khám phá vấn đề này bằng cách sử dụng các phương pháp**

**khác nhau.**

**Đối với Việt Nam, thị trường chứng khoán còn khá**

**mới mẻ, song ai cũng biết được tiềm năng và lợi ích đáng.**

**kế của nó. Việc khai thác được thị trường này sẽ đem lại**

**Tợi ích kinh tế cao. Dự báo thị trường chứng khoán là một**

**công việc khá quan trọng để khai thác lĩnh vực này. Chính**

**vì vậy tôi đã chọn đề tài “Kỹ thuật khai phá dữ liệu chuỗi**

**thời gian áp dụng trong dự báo chứng khoán” đề làm**

**luận văn tốt nghiệp với mục đích hiểu được công nghệ**

**data mining cũng như ứng dụng to lớn của nó trong việc**

**dự báo, dự đoán xu hướng trong tương lai, đặc biệt là**

**trong lĩnh vực thị trường tài chính, chứng khoán từ đó có**

**những quyết định đầu tư, giao địch phù hợp.**

**2. Mục đích nghiên cứu.**

**~\_ Nghiên cứu khái niệm, vai trò, ứng dụng và các kỹ**

**thuật khai phá dữ liệu.**

**- \_ Tìm hiểu kỹ thuật phân tích dữ liệu chuỗi thời gian**

**trong khai phá dữ liệu áp dụng vào bài toán dự báo nói**

**chung và dự báo trong thị trường chứng khoán nói riêng.**

**- Tìm hiểu mô hình ARIMA (Auto Regressive**

**Tntegrate Moving Average) với chức năng nhận dạng**

**mô hình, ước lượng các tham số và đưa ra kết quả dự báo.**

**dựa trên các tham số ước lượng đã được lựa chọn một**

**cách tối ưu. Thực nghiệm mô hình ARIMA trên dữ liệu**

**thời gian thực, áp dụng với dữ liệu chứng khoán hướng tới**

**việc dự báo chứng khoán.**

**3. Đối tượng và phạm vi nghiên cứu**

**Nghiên cứu các kỹ thuật khai phá dữ liệu, tập trung.**

**vào kỹ thuật phân tích chuỗi theo thời gian áp dụng vào**

**bài toán dự báo sự lên xuống của thị trường chứng khoán.**

**Mô hình ARIMA thực nghiệm trên dữ liệu VNIndex,**

**ABT, ACB.**

**4. Phương pháp nghiên cứu**

**Nghiên cứu, tìm hiểu lý thuyết về các kỹ thuật khai**

**phá dữ liệu.**

**Tìm hiểu, phân tích dữ liệu tài chính, chứng khoán.**

**Tìm hiểu cơ sở lý thuyết về mô hình ARIMA cho dữ**

**liệu thời gian thực (me series) và cách áp dụng vào bài**

**toán thực tế - dự báo sự lên xuống của thị trường chứng**

**khoán.**

**Xây dựng và thi hành mô hình ARIMA và ứng dụng**

**vào bài toán khai phá dữ liệu chị**

**thời gian trong dự báo**

**tài chính, chứng khoán**

**Sử dụng phần mềm Eviews đề thỉ hành chương trình.**

**Đánh giá kết quả dự báo được.**

**5. Kết cấu luận văn**

**Nội dung chính của luận văn chia làm 3 chương:**

**Chương 1: Tổng quan về khai phá dữ liệu giới thiệu**

**tổng quan về quá trình phát hiện tri thức và khai phá dữ**

**liệu, các kỹ thuật khai phá dữ liệu và ứng dụng của khai**

**phá dữ liệu.**

**Chương 2: Kỹ thuật khai phá dữ liệu chuỗi thời**

**gian giới thiệu về dữ liệu chuỗi thời gian thực và bài toán**

**dự báo đang được quan tâm trong khai phá dữ liệu. Giới**

**thiệu cơ sở lý thuyết của mô hình ARIMA và các bước**

**phát triển mô hình. Bài toán dự báo được áp dụng dưới**

**khía cạnh sử dụng mô hình ARIMA cho chuỗi thời gian**

**thực. Tiếp đến giới thiệu về phần mềm Eviews cho quá.**

**trình thi hành.**

**Chương 3: Ấp dụng mô hình ARIMA cho bài toán**

**dự báo chứng khoán trình bày thực nghiệm bài toán dự**

**báo với chuỗi dữ liệu tài chính, chứng khoán bằng mô**

**hình ARIMA. Thi hành các bước trong mô hình với phần**

**mềm Eviews 6, đưa ra kết quả và đánh giá với thực tế.**

**Cuối cùng là Phần kết luận và hướng phát triển của.**

**đề tài.**

**Chương 1: TỎNG QUAN VỀ KHAI PHÁ DỮ LIỆU**

**1.1.Giới thiệu**

**1.1.1. Khái niệm**

**+ \_ Khai phá dễ Tiệu (Data Mining)**

**=\_ Khám phá trí thức (Knownledge Discovery - KD)**

**Data Mining là một quá trình trích xuất thông tin có**

**mối quan hệ hoặc có mối tương quan nhất định từ một kho.**

**dữ liệu lớn (cực lớn) nhằm mục đích dự đoán các xu thế,**

**các hành vi trong tương lai, hoặc tìm kiếm những tập**

**thông tin hữu ích mà bình thường không thể nhận diện**

**được.**

**1.L2.Qu4 trình phát hiện trì thức trong CSDL**

**Mẫu đánh giá**

**Khai phá diện**

**"Diệp liên quan s“**

**x**

**Khodh hậu -—\_ ⁄/dhep**

**Tâm sịh dờliện**

**'Hình 1.1. Quá trình phát hiện tri thức.**

**1.2. Các kỹ thuật khai phá đữ liệu**

**1.2.1. Cây quyết định**

**1.2.2. Mạng nơron**

**1.2.3. Phân cụm**

**1.2.4. Luật kết hợp**

**1.2.5. Factor analysis (Phân tích nhân tô)**

**1.2.6. Chuỗi thời gian**

**1.3, Ứng dụng của khai phá dữ liệu**

**1.3.1. Dạng dữ liệu có thể khai phá**

**Data Mining được ứng dụng rộng rãi nên nó có thể**

**làm việc với rất nhiều kiểu dữ liệu khác nhau, một số dạng.**

**dữ liệu điển hình như: CSDL quan hệ, CSDL đa chiều**

**(mulidimentonal struetures, data warehouses), CSDL**

**dạng giao dịch, CSDL quan hệ-hướng đối tượng, dữ liệu**

**không gian và thời gian, Dữ liệu chuỗi thời gian, CSDL đa**

**phương tiện, dữ liệu Text và Web.**

**1.3.2. Ứng dụng của khai phá dữ liệu**

**Khai phá dữ liệu là một lĩnh vực được quan tâm và**

**ứng dụng rộng rãi. Một số ứng dụng điển hình trong khai**

**phá dữ liệu có thể liệt kê: (¡) phân tích dữ liệu và hỗ trợ ra**

**quyết định; đi) điều trị y học; đi) phát hiện văn bản; (iv)**

**tin sinh học; (v) tài chính và thị trường chứng khoán; (vi)**

**bảo hiểm...**

**1.3.3.Úng dụng của các kỹ thuật KPDL trong thị trường.**

**chứng khoán**

**Ứng dụng điển hình của khai phá dữ liệu trong thị**

**trường tài chính, chứng khoán đó là: phân tích tình hình**

**tài chính và dự báo giá của các loại cỗ phiếu trong thị**

**trường chứng khoán từ đó mang lại cho các nhà đầu tư**

**nhiều cơ hội để chọn lựa loại cỗ phiếu cần đầu tư, có hình.**

**thức và quy mô giao dịch phù hợp nhằm đạt được giá trị**

**gia tăng hiệu quả.**

**143.3.1. Ứng dụng của cây quyết định**

**1.3.3.2. Ứng dụng của mạng nơron**

**1.3.3.3. 'Ứng dụng của phân cụm**

**143.3.4. 'Ứng dụng của luật kết hợp**

**143.3.5. 'Ứng dụng của phân tích nhân tố**

**1.3.3.6. 'Ứng dụng của time series**

**Chương 2: KỸ THUẬT KHAI PHÁ DỮ LIỆU CHUÔI**

**THỜI GIAN**

**2.1. Bài toán dự báo**

**Dự báo là một nhu cầu không thể thiếu cho những**

**hoạt động của con người trong bối cảnh bùng nỗ thông tin.**

**Dự báo sẽ cung cấp những cơ sở cần thiết cho các hoạch.**

**định, và có thể nói rằng nếu không có khoa học dự báo thì**

**những dự định tương lai của con người vạch ra sẽ không**

**có sự thuyết phục đáng kể.**

**Có rất nhiều phương pháp, kỹ thuật để giải quyết bài**

**toán dự báo, trong đó có phương pháp dự báo theo chuỗi**

**thời gian. ARIMA là mô hình dự báo định lượng theo thời**

**số dự báo sẽ phụ thuộc vào**

**gian, giá trị tương lai của biết**

**xu thế vận động của đối tượng đó trong quá khứ (chuỗi dữ**

**liệu quá khứ).**

**2.2. Dữ liệu chuỗi thời gian**

**Một chuỗi thời gian (Time Series) là một chuỗi các.**

**quan sát theo trật tự thời gian. Chủ yếu những quan sát**

**này được thu thập ở những khoảng thời gian rời rạc, cách**

**đều nhau. Các mô hình chuỗi thời gian được đặc biệt áp**

**dụng trong dự báo ngắn hạn. Trong các bài toán dự báo**

**nói chung và các bài toán dự báo tài chính và chứng khoán**

**nói riêng, dữ liệu thường được biểu diễn dưới dạng chuỗi**

**thời gian. Trong các dạng dữ liệu được phân tích thì dữ**

**liệu chuỗi thời gian luôn thuộc tốp đầu về tính ph biến.**

**2.2.1. Chuỗi thời gian thực.**

**2.2.2. Thành phần xu hướng dài hạn**

**2.2.3. Thành phần mùa**

**2.2.4. Thành phần chu kỳ**

**2.2.5. Thành phần bắt thường**

**2.3. Mô hình ARIMA cho dữ liệu chuỗi thời gian**

**2.3.1. Các công cụ áp dụng trong mô hình**

**2.3.1.1. Hầm tự tương quan ACF (AutoCorrelation**

**Function)**

**XIÈƠi— HO —H)**

**\_ễŠẰằ;,ẰẶ7<.Ắ. @.Ð**

**2.3.1.2. Hàm tự tương quan từng phần PACE**

**T.=**

**3(f+k) = Cụp(tk-1) + Cay(t+k-2) + ... + Cụ-py(£ # 1) +**

**Cuo() + e(t) (2)**

**Tổng quan, hàm tự tương quan từng phần được tính**

**theo Durbin :**

**.¡ a9**

**2.3.1.3. Mô hình AR(p)**

**}(f)=agt+a¡y(f-1)+4ay(1-2)+...4„y(f-p)+e(f) (2.4)**

**Mô hình AR()): y(4) = # + a,y(f-1) + e(t)**

**Mô hình AR(): y() = áp + a,y(f-]) + ay(-2) +e(t)**

**2.3.1.4. Mô hình MA(q)**

**0) =bạ + e(0) +b¡£(t-1) + b;£(-2) +... +b£(-4) - (2⁄8)**

**Mô hình MA) : y() = bạ + e(£) + bue(†-1)**

**Mô hình MAO) : y() = bạ + £(t) + bu£(t-1) + b;£((-2)**

**2.3.1.5. Sai phân I(q)**

**Sai phân lần 1 (1) : zí?) = y(4)~y(-1)**

**Sai phân lần 2 (2) : hít) = z(t)~ z(-1)**

**2.3.2. Mô hình ARIMA**

**~ Mô hình ARMA(p,q):**

**3) = aptary(t-1)+aay(†-2)+...+auy(t-p)+e(1)**

**+bie(f-1)+bze(t-2)+...+bạe(-4) G6**

**~ Mô hình ARIMA(p,d,q):**

**Mô hình ARIMA (1, 1, 1):**

**Œ)~y(-1) = ap+ ai(y(t-])~ y(-2) + e(f) + bie(f-1))**

**Hoặc z(1) = ao + aiz(t-1) + e() + biế(t-1),**

**Với s(t) = y(t)— y(-1) ở sai phân đầu tiên: d**

**Tương tự ARIMA (1,2,1):**

**h(t) = aạ + aiz(t-1) + e(t) + b¡e(†-1),**

**Với hít) = z(/)— z(f-1) ở sai phân thứ hai: đ= 2.**

**2.3.3. Các bước phát triển mô hình.**

**2.3.3.1. Xác định mô hình**

**2.3.3.2. Ước lượng tham số**

**2.3.3.3. Kiểm định độ chính xác**

**2.3.3.4. Dự báo**

**Tĩnh các hàm tự tương quan và tự**

**tương quan riêng phẫn để nhận**

**dạng một mô hình dự định**

**——> Chọn lựa một mô hình.**

**Ước lượng các giá trị cho các,**

**tham số mô hình**

**Không**

**Kiểm định độ chính xác**

**của mô hình ước lượng.**

**Có**

**Sử dụng mô hình để dự báo**

**'Hình 2.16. Sơ đỗ mô phöng mô hình Box - Jenkins**

**2.4. Phần mềm EVIEWS**

**2.4.1. Giới thiệu phần mềm ứng dụng Eviews**

**Wosrea 'Siduelne**

**Hình 2.17.Cửa sỗ chính của Eviews [Nguồn: Eviewsr**

**5 Users Guide, tr16]**

**2.4.2. Áp dụng Eviews thi hành các bước của mô hình**

**ARIMA**

**2.4.2.1. Xác định mô hình**

**2.4.2.2. Ước lượng mô hình, kiểm tra mô hình**

**2.4.2.3. Dự báo**

**Chương 3: ÁP DỤNG MÔ HÌNH ARIMA CHO BÀI**

**TOÁN DỰ BÁO CHỨNG KHOÁN.**

**3.1. Dữ liệu tài chính, chứng khoán**

**Dữ liệu chứng khoán được biết tới như một chuỗi thời**

**gian đa dạng bởi có nhiều thuộc tính cùng được ghỉ tại**

**một thời điểm nào đó. Các thuộc tính của dữ liệu chứng.**

**khoán đó là: Open, High, Low, Close, Volume**

**3.2. Mô hình ARIMA cho dự báo chứng khoán**

**3.2.1. Quá trình xây dựng mô hình**

**- Xác định mô hình**

**~ Uúc lượng, kiểm tra mô hình**

**- Dự báo**

**3.2.2. Thiất kế mô hình ARIMA cho dữ liệu**

**Các bước để xây dựng một mô hình như sau :**

**1. Chọn tham biến**

**2. Chuẩn bị dữ liệu.**

**+ Xác định tính dừng của chuỗi dữ liệu**

**+ Xác định yếu tố mùa vụ**

**+ Xác định yếu tố xu thế**

**3. Xác định các thành phần p, q trong mô hình ARMA.**

**4. Ước lượng các nể**

**hợp nhất**

**ố và chẩn đoán mô hình phù**

**5. Dự báo ngắn hạn**

**3.3. Thực nghiệm**

**Sử dụng mô hình ARIMA và phương pháp Box —**

**Tenkins để thực hiện 3 quá trình dự báo giá đóng cửa của:**

**VnIndex, mã cổ phiếu ABT (của Công ty cổ phần xuất**

**nhập khẩu thủy sản Bến Tre) và mã cổ phiếu ACB (của**

**Ngân hàng Thương mại cổ phần Á Châu) trong ngắn hạn**

**căn cứ vào các chuỗi dữ liệu quá khứ của các mã CK đó.**

**3.2.1. Môi trường thực nghiệm**

**3.2.2. Dữ liệu đầu vào**

**Dữ liệu đầu vào của luận văn được lấy từ**

**hulp://www.cophieu68.com/datametastock.php. Đó là 3**

**file.CSV tương ứng với 3 mã CK được lấy từ website trên**

**xuống. Dữ liệu có dạng:**

**s= = [wac€———Ngay— SGsDengeus**

**[lAwoox 16/0A mo |AST —- M0202 „**

**[ưunoo son szs| |AED 9/0/05 +**

**[uunoer ago aa| |AD e0 ®**

**| uuoe œ/mJmn 2 sss| [am 0/022 xe**

**[ unoÐt oWum ai| [Am sen mà**

**kunoer sương — mg [Am aƯaớn mà**

**ưunos san su2| |asr màn"**

**mnoex sim 2 mi| |ar mm**

**| uuoex sưmion sss| [Am 2a**

**[ unoet z/ewAm su2| [Am zưemm sms**

**kAneer sự an soz| [Am 3saanp sẽ**

**unoex san sos| [Am miannn — sơ**

**kuuwmer z/ew2mp wo2| |Ar z/ann sex**

**[ưunoo 2/2 meớ| | 2ưa/nn — se**

**IETE——TSm- Tae**

**lộ nên**

**Ea | mẽ — mái**

**2S 22m nà**

**Em mm mới**

**la sưu — mới**

**2S 3n mà**

**-— m**

**PA. m**

**2o — Z/mzn> "**

**(¬ |Z mẽ — mà**

**2S 22 3)**

**-— ¬**

**mm...**

**š2 7mm — mà**

**'Hình 3.1. Dữ liệu đầu vào.**

**~>Tạo các workfile.**

**3.2.3. Xứ lý dữ liệu**

**3.2.3.1. Kiểm tra tính dừng của chuỗi chứng khoán**

**Dựa vào biểu đồ của biến giá đóng cửa của mỗi**

**chuỗi chứng khoán.**

**Hình 3.6. Biểu đồ giá đóng cửa của ABT**

**3.2.3.2. Nhận dạng mô hình**

**- Xác định các tham số p, d, q trong mô hình ARIMA.**

**của từng mã CK dựa vào biểu đồ tự tương quan.**

**Hình 3.9. Biểu đồ SAC và SPAC cũa chuỗi**

**GIADONGCUA của VNINDEX**

**3.2.3.3. Ước lượng và kiểm định với mô hình ARIMA.**

**=——¬**

**Tư mm DA tsgkrteonie SE <0i rõ**

**eedvesscz00ead[**

**Kemeneever**

**=——-...**

**re: 3000001 n6 n0.**

**Hình 3.16. Ước lượng mô hình ARIMA(1,0,1) cũa ABT**

**c3 ọioegel met becsliesel taneslreeeelsszlsecl**

**Tnhh 2g ấn Teen**

**em 582M2 p6Q0Hg xeaumo gian**

**NANU — c8mHƠ Cg8AEQ +eƯƠN giơ**

**CN. ':. sa.**

**Hình 3.17. Kết quả mô hình A RIMA(1,0,1) của ABT**

**Hình 3.18. Kiểm tra phần dư của chuỗi ABT.**

**Bảng 3.2. Bảng tiêu chuẩn đánh giá các mô hình ARIMA.**

**của ABT**

**Mô hình BIC Adjusted RẺ SEE**

**ARIMA**

**ARIMA(L00) | 2385271 0814950 - 0782972**

**ARIMA(J01) | 2345217 - 04825445 - 0.760445**

**ARIMA(,02) | 2397569 ' 0816063 ' 0780614**

**~3Mô hình được chọn cho chuỗi ABT là ARIMA(1,0,1)**

**3.2.3. Thực hiện dự báo.**

**Thực hiện dự báo giá đóng cửa của VNTNDEX, ABT,**

**ACB trong vòng 8 ngày từ 11/09/2012 đến 20/09/2012**

**Zaeastor**

**— tren nHaDIEcie**

**——- 3 Amence=eet**

**\_ƑẸ—A..**

**¬. g**

**suewa tơ sA)**

**G4120 ——.**

**nước**

**-**

**Needaen**

**-Y tết lờ eisfzrgkexev.ere**

**L7) La=)**

**Hình 3.23. Kết quä dự báo VNINDEX.**

**nh L Băng đánh giá giá dự báo VNINDEX so.**

**với giá thực tế**

**Ngày Giá dự báo | Giá thực | Đánh giá ¡ Sai số**

**tế (%)**

**11/09/2012 390.8433. 386.6. 4.2433 1.09**

**12/09/2012 391.1221 388.4. 27221 070**

**13/09/2012 ' 3913961 | 3914 | -00039 | ~000**

**14/09/2012 j 3916655 | 3989 | -72345 | 1485**

**17/09/2012 j 3919303 | 4018 | -9.8697 | 252**

**18/09/2012 j 3921906 | 3945 | -2.3094 | 0.59**

**19/09/2012 | 3924465 | 3946 | -21535 | 0455**

**20/09/2012 ; 3926980 | 3893 | 3.3980 | 0487**

**Đánh giá: kết quả dự báo là khá chính xác (mức độ sai**

**số rất thấp, từ xấp xi 0% đến 2.52%).**

**Luận văn đã trình bày được tổng quan về khai phá dữ**

**liệu: khái niệm, các kỹ thuật khai phá dữ liệu và các ứng**

**dụng của khai phá dữ liệu. Trong đó luận văn tập trung**

**vào kỹ thuật khai phá dữ liệu chuỗi thời gian áp dụng vào**

**bài toán thực tế đang được quan tâm đó là bài toán dự báo.**

**nói chung và dự báo giá chứng khoán nói riêng.**

**Luận văn cũng đã trình bày được một số nội dung cơ**

**sở lý thuyết về chuỗi thời gian thực, về mô hình ARIMA.**

**(các công cụ áp dụng trong mô hình, quy trình xây dựng**

**mô hình) và phần mềm Eviews, áp dụng Eviews để thỉ**

**hành các bước của mô hình ARIMA trong dự báo chứng**

**khoán. Tác giả cơ bản nắm được quy trình dùng phần**

**mềm Eviews để xây dựng mô hình ARIMA cho dữ liệu**

**thời gian thực, tính toán giá trị dự báo cho chuỗi dữ**

**chứng khoán.**

**Luận văn đã áp dụng những cơ sở lý thuyết nghiên**

**cứu tiến hành thực nghiệm trên ba chuỗi chứng khoán (chỉ**

**số Vnlndex, mã CK ABT, ACB) dựa trên dữ liệu lịch sử**

**của mỗi chuỗi (gồm 257 quan sát trong quá khứ) và đã dự**

**báo được giá đóng cửa của 10 ngày tiếp theo. Kết quả dự**